From Human Routine to More Efficient Mobile Networks

Eduardo Mucelli Rezende Oliveira

DE LA ROUTINE HUMAINE VERS DES RÉSEAUX MOBILES PLUS EFFICACES

EDUARDO MUCELLI REZENDE OLIVEIRA

Laboratoire d'Informatique

Thèse présentée pour obtenir le grade de Docteur en Informatique de L'École Polytechnique. Soutenue publiquement le 11 Mai 2015 devant le jury composé de :

Aline CARNEIRO VIANA	INRIA Saclay, France	(Directrice de thèse)
Marcelo D. DE AMORIM	CNRS/Université Pierre et Marie CURIE, France	(Rapporteur)
Marco FIORE	National Research Council, Italie	(Rapporteur)
Thierry TURLETTI	INRIA Sophia Antipolis, France	(Examinateur)
Philippe JACQUET	Bell Labs et École Polytechnique, France	(Examinateur)
Jean-Marie GORCE	INSA Lyon, France	(Examinateur)

DIRECTRICE DE THÈSE:
Aline Carneiro Viana

LIEU:
Palaiseau, France

FROM HUMAN ROUTINE TO MORE EFFICIENT MOBILE NETWORKS

EDUARDO MUCELLI REZENDE OLIVEIRA

Laboratory of Informatics

Thesis presented to obtain the grade of Doctor in Computer Science of the École Polytechnique. Publicly defended on 11 May, 2015 in the presence of the jury composed by:

Aline CARNEIRO VIANA	INRIA Saclay, France	(Advisor)
Marcelo D. DE AMORIM	CNRS/Université Pierre et Marie CURIE, France	(Reviewer)
Marco FIORE	National Research Council, Italy	(Reviewer)
Thierry TURLETTI	INRIA Sophia Antipolis, France	(Examiner)
Philippe JACQUET	Bell Labs and École Polytechnique, France	(Examiner)
Jean-Marie GORCE	INSA Lyon, France	(Examiner)

SUPERVISOR:
Aline Carneiro Viana

LOCATION:
Palaiseau, France

RÉSUME

L'omniprésence des communications a entraîné une récente augmentation des volumes de données mobiles, pour laquelle les opérateurs n'étaient pas toujours préparés. Les smartphones sont les plus gros consommateurs de données mobiles. Ces appareils peuvent être considérés comme méchants à cause d'un tel traffic, mais d'un point de vue analytique ils fournissent, aujourd'hui un des meilleurs moyens afin de collecter les données sur le comportement de consommation et de mobilité de grande échelle. Comprendre le comportement des utilisateurs sur leur mobilité et leur connectivité est nécessaire à la création d'un système de communication effectifs. Nous sommes routiniers. Ces cycles routiniers sont une grande partie de nos interactions avec le monde. Par exemple, nos habitudes definissent ce que l'on va faire le samedi ou les sites que nous consultons le lundi matin. Ces comportements répétés reflètent nos déplacements et activités en ligne. Dans cette thèse, nous allons nous concentrer sur les demandes de traffic générées par les usagers métropolitains durant leurs activités quotidiennes. Nous présentons une étude détaillée des usagers selon les comportements routiniers de mobilité ou d'activité sur internet. Dans une étude de cas, ou cette enquête serait utile, nous proposons une stratégies de déploiement de points de accès qui prendra en compte les aspects routiniers de la mobilités des utilisateurs.

Nous étudirons en premier lieu, les modèles de mobilité en milieu urbain. Nous analyserons les données de mobilité à grande échelle dans de grandes villes comme Beijing, Tokyo, New York, Paris, San Francisco, London, Moscow, Mexico City. Cette contribution se fait en deux étapes. Premièrement, nous observerons les similitudes des déplacements peu importe la ville concernée. Ensuite, nous mettrons en évidence trois caractéristiques présentes dans les déplacements d'une population urbaine typique: Répétivité, utilisation de raccourcis, confinement. Ces caractéristiques sont dues à la tendance qu'ont les personnes à revisiter les même rues en utilisant les trajectoires proches du chemin le plus court. D'ailleurs, les personnes ont une mobilité quotidienne inférieure à dix kilomètres par jour.

Nous avons ensuite étudié les modèles de demandes de traffic en utilisant une base de données comprenant les données de 6.8 millions d'utilisateurs. Pour cela nous avons principalement deux contributions. Premièrement, une caractérisation précise des comportements de consommation des utilisateurs agrégés par modèle. Nous pouvons voir comment les routines quotidiennes impactent nos demandes de connections et la similarité de ce traffic en fonction des jours. Ensuite, nous fournirons un moyen de reproduire artificiellement mais

avec cohérence les modèles des utilisateurs de données mobiles. Ces données synthétisées ont l'avantage de permettre la planification du réseau sans information sur la vie privées de utilisateurs comme les bases de données d'origine.

Afin d'évaluer l'efficacité de ces informations dans un scénario grandeur nature, nous proposerons une stratégie de deploiement de points de accès qui prend en compte les caractéristiques routinières en terme de déplacement et de demande de trafic dans le but d'améliorer la décharge de données mobile. Déployer correctement des points de accès WiFi peut être moins cher que d'améliorer l'infrastructure de réseaux mobiles, et peut permettre d'améliorer considérablement la capacité du réseau. Notre approche améliore l'évacuation de trafic comparée aux autres solutions disponibles dans la littérature.

ABSTRACT

The proliferation of pervasive communication caused a recent boost up on the mobile data usage, which network operators are not always prepared for. The main origin of the mobile network demands are smartphone devices. From the network side those devices may be seen as villains for imposing an enormous traffic, but from the analytical point of view they provide today the best means of gathering users information about content consumption and mobility behavior on a large scale. Understanding users' mobility and network behavior is essential in the design of efficient communication systems. We are routinary beings. The routine cycles on our daily lives are an essential part of our interface with the world. Our habits define, for instance, where we are going Saturday night, or what is the typical website for the mornings of Monday. The repetitive behavior reflects on our mobility patterns and network activities. In this thesis we focus on metropolitan users generating traffic demands during their normal daily lives. We present a detailed study on both users' routinary mobility and routinary network behavior. As a study of case where such investigation can be useful, we propose a hotspot deployment strategy that takes into account the routine aspects of people's mobility.

We first investigate urban mobility patterns. We analyze large-scale datasets of mobility in different cities of the world, namely Beijing, Tokyo, New York, Paris, San Francisco, London, Moscow and Mexico City. Our contribution is this area is two-fold. First, we show that there is a similarity on people's mobility behavior regardless the city. Second, we unveil three characteristics present on the mobility of typical urban population: repetitiveness, usage of shortest-paths, and confinement. Those characteristics undercover people's tendency to revisit a small portion of favorite venues using trajectories that are

close to the shortest-path. Furthermore, people generally have their mobility restrict to a dozen of kilometers per day.

We then investigate the users' traffic demands patterns. We analyze a large data set with 6.8 million subscribers. We have mainly two contributions in this aspect. First, a precise characterization of individual subscribers' traffic behavior clustered by their usage patterns. We see how the daily routine impacts on the network demands and the strong similarity between traffic on different days. Second, we provide a way for synthetically, still consistently, reproducing usage patterns of mobile subscribers. Synthetic traces offer positive implications for network planning and carry no privacy issues to subscribers as the original datasets.

To assess the effectiveness of these findings on real-life scenario, we propose a hotspot deployment strategy that considers routine characteristics of mobility and traffic in order to improve mobile data offloading. Carefully deploying Wi-Fi hotspots can both be cheaper than upgrade the current cellular network structure and can concede significant improvement in the network capacity. Our approach increases the amount of offload when compared to other solution from the literature.

PUBLICATIONS

This thesis resulted in the following list of publications:

PUBLISHED

- Eduardo Mucelli Rezende Oliveira, Aline Carneiro Viana, K. P. Naveen and Carlos Sarraute, "Analysis and Modeling of Mobile Data Traffic in Mexico City", *NetMob Conference*, April 2015, Mit Media Lab, Cambridge, MA, United States.
- Eduardo Mucelli Rezende Oliveira, Aline Carneiro Viana, K. P. Naveen and Carlos Sarraute, "Measurement-driven mobile data traffic modeling in a large metropolitan area", *IEEE Percom* 2015, Saint Louis, United States.
- Eduardo Mucelli Rezende Oliveira and Aline Carneiro Viana, "From Routine to Network Deployment for Data Offloading in Metropolitan Areas", *IEEE SECON*, June 2014, Singapore.
- Eduardo Mucelli Rezende Oliveira and Aline Carneiro Viana, "From Routine to Network Deployment for Data Offloading in Metropolitan Areas", *École d'eté RESCOM*, December 2013, Lyon, France.
- Eduardo Mucelli Rezende Oliveira and Aline Carneiro Viana, "Routine-based Network Deployment", *IEEE INFOCOM*, Student workshop, April 2014, Toronto, Canada.
- Eduardo Mucelli Rezende Oliveira and Aline Carneiro Viana, "From Routine To Better Network Services", *IEEE/IFIP WMNC*, May 2014, Algarve, Portugal.
- Eduardo Mucelli Rezende Oliveira and Aline Carneiro Viana, "Routine-based network deployment for data offloading in metropolitan areas", *IEEE WCNC* 2014, Istanbul, Turkey.
- Eduardo Mucelli Rezende Oliveira and Aline Carneiro Viana, "From your routine to hotspot deployment for data offloading", *ACM CoNEXT* Student workshop 2012, Nice, France.

UNDER REVIEW

- Eduardo Mucelli Rezende Oliveira, Aline Carneiro Viana, K. P. Naveen and Carlos Sarraute, "Mobile Data Traffic Modeling: Re-

vealing Temporal Facets", *IEEE Transactions on Mobile Computing*.

- Eduardo Mucelli Rezende Oliveira, Aline Carneiro Viana, K. P. Naveen, Carlos Sarraute, Jorge Brea, and Ignacio Alvarez-Hamelin, "On the Consistency of Human Mobility Analysis", *Elsevier Pervasive and Mobile Computing*.

ACKNOWLEDGMENTS

"All that I am, or hope to be, I owe to my angel mother."

— Abraham Lincoln

Minha mãe, a pessoa mais importante na minha vida. Obrigado por estar sempre comigo, mesmo que em distâcia, sempre me acompanhar, me incentivar e me compreender.

"A happy family is but an earlier heaven."

— George Bernard Shaw

À minha família por ter dado o apoio fundamental e necessário para que hoje eu possa concluir este trabalho. Mesmo que a distância seja grande, sinto abraçá-los todos os dias. Agradeço a minha irmã pelas conversas, pelo apoio e pela amizade, valeu, Isabele *"Bibilz"* Mucelli R. de Oliveira. Ao meu pai, por sempre me motivar com sua sabedoria e jovialidade, valeu, José *"Gordon"* de Oliveira.

"Truly great friends are hard to find, difficult to leave, and impossible to forget."

— G. Randolf

To my friends, wherever you were, for the support, jokes, and good times spent together. Gustavo Dias da Silva, Felipe *"Big Head"* Domingos Cunha, K. P. Naveen, Kostas Tavlaridis-Gyparakis, Benoit Defossez, Thiago *"Pedpano"* Henrique Silva, and for my eternal friends that were far away, which I will never forget.

"Love is that condition in which the happiness of another person is essential to your own."

— Robert A. Heinlein

To my beloved Sandrine, for every single wonderful smile that your presence brings to me.

"A teacher affects eternity; one can never tell where his influence stops."

— Henry Brooks Adams

To all my teachers, specially to my advisor, for the necessary support and guidance in order to walk this long path until the accomplishment of this thesis work.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LISTINGS

ACRONYMS

1 INTRODUCTION

"The most profound technologies are those that disappear.
They weave themselves into the fabric of everyday life
until they are indistinguishable from it."

— Mark Weiser [1]

1.1 CONTEXT AND MOTIVATION

We are surrounded by pervasive devices, many of them connected to the Internet. They are growing in number and capacity. There will be 50 billion connected devices by 2020, i. e., more than 6 per person [2]. Smartphones are one of the most important instances of the connected devices. Smartphone sales surpassed feature phones for the first time in 2013 [3]. Almost 1 billion of them were sold to end users only in 2013, a growth of 42% when compared to the previous year [3]. From that, over half billion mobile devices and mobile network subscriptions were added in 2013. This is a noteworthy shift towards mobile web access. As a consequence, global mobile data traffic grew 83% in 2013 being 18 times the size of the entire global Internet in 2000 [4]. This is a steady trend, mobile data traffic will increase nearly 10-fold between 2014 and 2019. Figure 1 shows an interesting example of the rapid spread of mobile devices on two pictures taken 7 years apart during New Year's Eve In Rio de Janeiro's Copacabana Beach.

(a) (b)

Figure 1: Increase on the number of smartphone devices in 7 years, Rio de Janeiro's Copacabana beach (a) on 2008 and (b) on 2015 [1].

Using better web-enabled smartphones and under growing network coverage, users generate traffic every time and everywhere. For instance, per minute, people around the world upload 49,000 photos on Instagram [5], 244,000 photos on Facebook [6], 487,000 photos on

1 Ayrton 360 and Globo Television

WhatsApp [7], send 348,000 tweets [8], and upload 100 hours of video on YouTube [9].

In an interval of five years, from 2007 to 2012, AT&T faced 20,000 percent growth in mobile data traffic on its network [10]. Telco operators are struggling to cope with the rapidly growing capacity requirements whilst keeping financial health. To do so, some technical solutions are currently being implemented such as adoption of new communication standards (e. g., LTE), video optimization, and traffic shaping.

LTE has a cost efficient deployment when compared to technologies such as 2G and 3G for reasons such as part of its equipment can be collocated with existing ones, it has higher spectral efficiency and it demands less power consumption. Therefore, it costs less to deliver a byte using LTE than it does on 3G networks. Nonetheless, many years will be required until a significant adoption of 4G devices. On Western Europe, 25% to 30% of the data traffic will be consumed using 4G devices only by 2018 [11]. On a global perspective, 25% of the connections will be made by a 4G-enabled device only by 2020 [12].

Considering that by 2019, 72% of total mobile data traffic will be used for mobile video streaming, a 13-fold increase between 2014 and 2019 [4], Telco operators are investigating video optimization as one of the ways to face the unprecedented growth on network demands. The network load can be reduced, for example, by fine tuning the transcoding, transrating, or by lowering video resolution. However, the end-user experience may be seriously affected by the video quality degradation.

Since aforementioned approaches tackle the traffic inside the mobile network or try to couple with the traffic growth, there is a general belief that those solutions will not handle the demands on the next years. On the other hand, mobile traffic offloading seems to be a promising solution as it aims to shift the traffic off of the mobile network to Wi-Fi network. This approach has several advantages such as low cost, high data rates and easy availability of Wi-Fi hotspots. Careful deployment of Wi-Fi hotspots can both be cheaper than upgrade the current cellular network structure and can concede significant improvement in the network capacity [13, 14].

In order to plan a network it is important first to understand how its subscribers generate traffic. This is challenging task due to the differences on the behavior of each one of the users. Furthermore, planning a hotspot network require understanding of how people interact with the city, such as important places, which may be areas under high traffic. For instance, Telco operator's network on metropolitan areas may facilitate the collocation of Wi-Fi hotspots on important points of the city. Those hotspots would offload the data from subscribers passing close by, during their daily activities. People's mo-

bility and traffic behavior are tightly related to their daily routines. Thus, it is essential to understand the dynamics that guide the urban mobility and network traffic demands of people to better design network placement strategies.

The objective of this thesis is to get insights from mobile user's routinary behavior (in terms of context and content demands) for network performance improvement.

1.2 PROBLEM STATEMENT

In this thesis, we interrogate the current human mobility and network traffic analysis. We focus on two main questions and later, in a way to take advantage of them in the context of mobile networks.

1.2.1 *What about investigating routine instead of purely mobility?*

Considering the vast possibilities human beings have when it comes to their mobility, there is a considerable amount of aspects from which one can analyze trajectories. For instance, the mobility and the relation among people has been widely studied through wireless interaction analysis on intermittently connected networks (e. g., DTN, VANETs) networks. However, these studies overlook some key aspects inherent to human routinary behavior and the way he/she interacts with the environment. Understanding those aspects may positively impact on the efficiency of systems that rely on human mobility.

1.2.2 *Can we identify common traffic behavior among mobile subscribers?*

To better plan a mobile network it is crucial to understand its subscribers. That is a challenging task since each of the subscribers have their own traffic behavior, which seems to be unique among all other subscribers. Although a per-subscriber modeling is unpractical due to the large number of subscribers, there is an actual similarity in the network traffic demands among different subscribers, partly because of their routines. Thus, it is important to have a solution that identifies and agglutinates *similar* subscribers into profiles that represent common network behaviors. Additionally, once the traffic is modeled, it can be synthetically reproduced, which has advantages such as expanding the limit of the original data used in the model and the protection of subscribers' privacy.

1.2.3 *And so what?*

Once the users are better understood from their mobility and network point of views, we can envision benefits to both them and to the

network operator. Better quality of service, coverage, cost-benefit are among possible outcomes from the understanding of mobility and network traffic behavior. Our study case exemplifies a mobile data traffic offloading scenario, which takes advantage of the understanding from our users' behavior pattern analysis.

1.3 CONTRIBUTIONS OF THIS THESIS

1.3.1 *Routine characterization of human mobility*

When searching for patterns on the human mobility there is no evidence that a common set of characteristics exists for every significant sample of the population. The reason is that there is an enormous amount of possibilities one can take when roaming around during her/his daily activities. Since we are *free* to decide about our own mobility, one may expect a vast amount of variability on our daily trajectories. Although trajectory possibilities are immense, we are routinary beings with inclination to repeat cyclic activities. Literature presents some insights regarding human mobility for few specific cities.

As first contribution of this thesis, we study human mobility in order to find and measure a set of consistent characteristics present in the routine of human mobility. We analyze several different datasets of human mobility in 8 cities from 3 different continents. Moreover, we analyze mobility and its relation with urban scenario, i. e., points of interest in those cities. We see three core aspects present on the human mobility: the usage of shortest-paths, the confinement, and the repetitive visit to a set of points of interest.

1.3.2 *Mobile data traffic profiling and synthetic generation*

Network planning requires understanding of the traffic demands. Each subscriber has an important impact on the network traffic as a whole. A fine grained study of the network traffic requires a per-subscriber study, which is unpractical due to the singularity present on the behavior of each user.

The second contribution of this thesis consists on proposing a subscriber profiling methodology which define a finite set of classes based on network traffic behavior. We analyze a large-scale dataset with 6.8 millions subscribers in Mexico city on a period that spans for 4 months. Due to the cyclicity of our daily activities, we see that subscribers network demands present more similarity in the same hours on different days than within the same day on different hours. We take advantage of this periodic behavior to classify subscribers into 6 profiles using a subset of the dataset and considering two important parameters for the traffic generation: volume and frequency. Finally,

we propose a synthetic traffic generator that consistently imitates the network demands from the original dataset.

1.3.3 *Traffic-and-mobility-aware hotspot deployment for data offloading*

In order to apply the previous insights, we have chosen a hotspot deployment study case. Our approach takes into consideration a metropolitan area by leveraging mobile users' context and content, i. e., their trajectories, scenario interactions, and traffic demands. This is a challenging task due to the large area covered by big cities. For instance, literature presents complete coverage approaches for this problem on very limited scenarios, such as university campus. Such solutions are prohibitive in terms of cost when it comes to a large metropolitan area.

Our proposed strategy considers the restrictions imposed by transportation modes to people trajectories and the space-time interaction between people and urban locations, key points for an efficient network planning. Using a real-life metropolitan trace with 182 users and more than 200,000 points of interest in Beijing, we show our routine-based strategy guarantees higher offload ratio than the current approach in the literature, while using a realistic traffic model.

1.4 THESIS OUTLINE

The rest of this thesis is organized as follows. Chapter 2 presents a classification on the methods used to analyze mobility and network mobile traffic. In Chapter 3, we study several routine aspects in the human mobility. Then, in chapter 4, we propose a profiling methodology based on subscribers traffic demands and a synthetic traffic generator. Later, using previous insights, in Chapter 5, we propose a Wi-Fi hotspot deployment for mobile data offloading. Finally, Chapter 6 presents our conclusions and envisioned next steps.

2

ON THE MOBILITY AND CONTENT ANALYSIS

"We are like dwarfs on the shoulders of giants, so that we can see more than they, and things at a greater distance, not by virtue of any sharpness of sight on our part, or any physical distinction, but because we are carried high and raised up by their giant size."

— Bernard De Chartres

Related to the problems stated in the previous chapter, we present the bases of mobility and traffic dataset analysis as well as we position our work according to the literature.

2.1 DATASET KNOWLEDGE EXTRACTION

Datasets are of enormous importance on the analysis of many scientific fields. They provide the convenience of a non-real time analysis, i. e., one can analyze the phenomena of interest after its parameters have been collected and logged. In the context of large-scale mobility and networks, where real-time analysis is arduous due to the enormous amount of elements (e. g., subscribers) and parameters (e. g., position), datasets are widely used as primary source of information.

As a precious resource, few are the datasets in the literature that contain large-scale information of mobility or network measurements. Moreover, to the best of our knowledge, no freely available dataset has both large-scale information about fine-grained mobility and network traffic. Experiments to collect human mobility data generally involve people carrying GPS-capable devices which regularly collect their precise positioning. Due to the complexity of those experiments, they tend to be limited in number of participants (e. g., up to 35), time duration (i. e., up to few weeks), and space as in university campuses [15, 16, 17, 18, 19], conference rooms [20, 21], or shopping malls [22]. Lausanne campaign [23] and GeoLife [24] represent some of the few relatively large experiments with around 200 participants, that attempts to collect fine-grained human mobility. The dataset collected from the former is not publicly available, while the one from the latter is.

Aside from that, human mobility datasets covering large areas tend to rely *only* on automobile transportation, which is not in the scope of this thesis. For instance, taxi cabs in San Francisco [25] and Rome [26], vans inside Microsoft headquarters in Redmond [27], buses inside a university campus area [28] or in the metropolitan area of Seattle [29].

Another source of human mobility data is Call Detail Record (CDR). CDR is a metadata record that describes phone communication using

a series of data fields, e. g., the identification of callee and caller, call type (voice call or SMS), starting, ending and duration time of the call, and GPS location of the caller's cell tower [30]. CDR datasets are usually released by Telco operators to a limited number of partners under a non-disclosure agreement and with limited access. As both mobility and network traffic are susceptible of giving away private users' information, entities responsible for such data are careful on providing it to third-parties.

Besides, it is important to understand the limitations of each dataset. For instance, when modeling mobility through CDRs, one has to know its two biggest limitations: sparse in time and coarse in space. Time sparseness occurs because records are generated only when a subscriber sends or receive a call or a SMS, which makes he/she *invisible* at all other periods of time. Space coarseness is due to the granularity of a cell tower sector for the subscribers' positioning, which leads to a location uncertainty of about 1 square mile [31]. It is important to consider that those two characteristics are not uniformly distributed in time due to the fact that subscribers tend to place their calls in bursts, then staying nonactive for long periods [32], around 70% of the total time [31]. To overcome such limitation on mobility analysis, a threshold may be set aiming to remove from the dataset subscribers that have low call frequency [31]. Nonetheless, [33] studies the possible caveats of using CDRs to model human mobility patterns. It shows that this kind of data performs well in certain scenarios such as evaluation of subscribers' home and work places.

Similarly, network traffic datasets in large scale are rarely available in the literature. As in CDR datasets, they are generally released by Telco operators as xDRs, an extension of CDRs that includes network traffic usage information, e. g., generated from subscriber's browsing activities and background applications. In some cases, more specific information such every single URL recorded during subscribers browsing sessions are available as presented in [34].

Getting access to the dataset is the first step towards proposing a system model. Some effort has to be made to extract mobility and traffic knowledge from raw datasets due to their singular characteristics. For example, CDR mobility datasets generally contain, per user, a set of geolocalized points generally ascendingly sorted by timestamp per day. Assumptions must be made in order to represent disjoint trajectories that a user perform during the day, otherwise every user will have one single day-long trajectory, which might not represent accurately his displacement. Contrarily, for instance, GeoLife provides a precise set of trajectories per user, which ease the mobility analysis.

2.2 MOBILITY INSIGHTS

The understanding of mobility and its modeling has started with animals such as monkeys [35], jackals [36], and albatrosses [37]. Such works indicated that animal mobility follows a random walk for which the displacement is power-law distributed, i. e., Lévy flight [38]. Early human mobility studies used tracking methods such as bank notes dispersion [39]. Latterly, the lower cost of GPS devices increased the possibility of collecting mobility datasets. In [40], the authors evaluate GPS traces of 44 volunteers in various outdoor scenarios including two different college campuses, a metropolitan area, a theme park and a state fair. The analysis shows that human mobility resembles Lévy flight within a scale of less than 10 km, which corroborates the findings from [39]. Authors then create a Lévy flight model that captures the mobility from those individuals. More recently, easier methods for collecting human mobility in large scale such as mobile phones open new horizons for deeper human mobility investigations.

Through extensive analysis, [41] presents a seminal study on human mobility using a CDR dataset of 100,000 subscribers. Authors show that human trajectories show a high degree of temporal and spatial regularity, in disagreement with the aforementioned random trajectories predicted by the prevailing Lévy flight and random walk models. Besides, each individual is characterized by a specific travel distance that is time-independent and a significant probability to return to a few highly frequented locations. The return to a previously visited location occurs with a frequency proportional to the ranking in popularity of the location with respect to other locations. It means, that humans have a strong tendency to return to locations that they visited before, due to the recurrence and temporal periodicity inherent to human mobility. [42] presents an extension of this work using two CDR datasets totalizing 3 million subscribers focusing on the visiting time, i. e., the period of time spent at one location. The resulting curve shows a truncated power-law with a cutoff of 17 hours, which authors link to the typical awake period of humans.

[43] analyzes CDRs dataset of 97,000 subscribers in Los Angeles and 71,000 in New York aiming to identify important locations in peoples' lives. Using ground-truth data of home and work location from 19 subscribers, authors were able to identify home and work locations with about 1 and 21 miles of error, respectively. In [44] authors evaluate a dataset of CDRs with information of about 450,000 subscribers to capture city dynamics. More specifically, authors want to discover two main groups in a city, one active during the day (*laborshed*) and another during the night (*partyshed*). Their grouping strategy relies on a set of fixed rules, e. g., a subscriber is set to laborshed group if he makes 4 calls (or send 4 SMSs) during business hours using city cell towers, at least, twice per week. It is shown a 81% correlation between

the groups identified by their algorithm and US Census dataset used as ground-truth.

In [45] authors analyze a subset of Lausanne dataset [23] with 38 participants in order to understand how temporal and personal factors, e. g., occupation and age, affect individual mobility patterns. From the temporal analysis, they concluded that people are less active during workdays and night than during weekends and daytime. Occupational analysis shows that among full-workers and students, the former are more prone to shorter displacement during the day due to the stricter time rules imposed on companies when compared to universities. Finally, age analysis shows higher nightly mobility of younger people compared to older counterparts, which is the result of nightlife attractions being more interesting for younger people. In [46], a study was made using a CDR dataset containing information of 180 subscribers, which presented similar temporal findings.

Due to the routinary behavior, human mobility is highly predictable. [31] presents a study using a CDR dataset with 50,000 subscribers aiming to measure how predictable human mobility is. Authors measure the predictability of subscribers' next whereabouts by using three entropy metrics: (1) uniformly randomly chosen among all the locations the subscriber already visited, (2) based on the frequency of the most visited locations, or (3) taking into consideration both frequency, time spent and the order of the visits. As result, for the typical subscriber, the uncertainty of the next location (i. e., the cell tower the subscriber will be connected to) resides, on average, in a set of less than two locations. Moreover, [41] shows that individuals are found at their first two preferred locations on 40% of the time.

Besides the efforts above, human mobility has been widely studied from several points of view, specially with regards to the inter-contact and contact time between people, i. e., the time gap separating two contacts and its duration considering the same pair of people. The importance on those studies comes from a specific problem on intermittently connected networks: as messages are transmitted among nodes when they get in contact with each other, the contact time between pairs of nodes is a key factor on the end-to-end communication delay. In the context of human mobility, people carrying mobile phones are nodes and a contact between devices signify respective people getting closer to each other. The longer they stay close, i. e., the contact duration, the larger the amount of data that can be exchanged.

In [47, 48] authors show that empirical distributions of inter-contact times present two characteristics. First, they are well fitted by log-normal curves, with exponential curves also fitting a significant portion of the distributions. Second, they can be well approximated by a power law over some specific time ranges, from few minutes to 12 hours. [21, 49, 50] conducted experiments involving Bluetooth contacts between people carrying devices: [21] studies data from 41 par-

ticipants at Infocom 2005 conference rooms, [49] analyses 9 participants in a campus scenario, and [50] assess data collected from 16 undergraduate students. Similar results are present in those works regarding contact duration: it is power-law distributed with variations in the power-law coefficient k inherent to the specificities on the scenarios where the experiments were carried-out. For instance, the contact duration distribution curve presented in [50] decays slower when compared to the ones from [17, 49]. Authors associate this behavior to students that tend to stay longer periods of time in the vicinity of each other as they may attend to the same classes.

Aforementioned works have mostly studied some aspects of the human mobility unveiling characteristics on people's displacement such as distance, high probability to revisit certain few locations, and the dynamic of encounters. Their conclusions indicate that temporal and spatial factors are recurrently impacting human mobility. However, our intuition says that people's mobility presents other characteristics such as tendency to use shortest-paths. Besides, no large scale evaluation of fine-grained datasets was performed to verify this intuition nor the aspects previously assessed in the literature.

2.3 DATA TRAFFIC INSIGHTS

The understanding of users' content consumption has attracted significant attention of the networking community in the literature. Its improved understanding is of fundamental importance when looking for *solutions to manage the increased data usage and to improve the quality of communication service provided*. The resulting knowledge can help to design more adaptable networking protocols or services, as well as to determine, for instance, where to deploy networking infrastructure, how to reduce traffic congestion, or how to fill the gap between the capacity granted by the infrastructure technology and the traffic load generated by mobile users.

Earliest analysis of cellular network traffic were mostly focused on the traffic generated from micro-browsing using Wireless Application Protocol (WAP) [51], PDAs [52], and CDMA2000 network technology [53]. Although those studies were made more than a decade ago, certain findings are still somehow similar to the current state of the literature, e. g., in [51] authors detected day cyclicity on the micro-browsing access behavior of mobile phone users. Besides, [52] shows that most of the users tend to have short network usage sessions when accessing web sites. Compared to those first studies, current network traffic data collection and analysis involves an enormous amount of data, which raises the difficulty on the understanding, characterization and classification of network demands and subscribers.

Within network traffic analysis, considerable effort has been made towards the classification of subscribers behavior based on their sim-

ilarities. This is challenging task due to the heterogeneity of human behavior, while some subscribers rarely make use of the voice calls, others perform them thousands of times per month [54]. Clustering techniques are of frequent use in this context and are employed since the earliest analysis on metropolitan-area wireless network traffic [55]. Clustering primary focus on the concept of similarity between elements in a collection, as determined by the distance between them in a multidimensional space. Two elements belong to the same cluster if the distance between them is small enough. Relatively simple clustering algorithms have the capability to group a large number of elements into clusters, which generally are predefined classes with a semantical meaning.

Distinguishing between changes that are merely due to the dynamic nature of the system and anomalies is a difficult problem. In [56] authors develop a method to identify anomalous behavior from subscribers' call records. They apply a clustering technique to detect anomalous voice call pattern relative to the historical pattern established for a phone. Their evaluation makes usage of a CDR dataset comprising 500 subscribers. Their approach uses artificial neural networks model and focus only on the voice call duration parameter. A voice call is considered abnormal if it is way longer than the common subscriber's voice call length. Similarly, [57] applies clustering techniques in order to find anomalous network behavior. This approach combines Leader [58] and K-means [59] clustering algorithms into a *hybrid* scheme that considers voice call duration and starting time, number of sent/received SMSs, and data traffic. Authors evaluate the framework on a 12-day CDR dataset of unspecified amount of subscribers. Resulting clusters with few members were considered to contain records anomalous voice calls. Authors evaluate several configuration for the clustering algorithms parameters and conclude that an hybrid approach composed of Leader and K-means have very limited success on anomaly detection.

In [60] authors use a CDR dataset with information about 475,000 subscribers aiming to classify subscribers into usage groups. In this study, voice call duration and number of SMSs are the only parameters that represent subscriber's *usage*. In order to create the groups, K-means clustering was used on both parameters and $k = 7$, i. e., a fixed number of seven usage groups. Moreover, further investigation in one of the clusters show predominant activity before and after working hours, which authors link to a cluster with majority of commuter subscribers. In another cluster, authors found higher number of SMSs than calls, which is a characteristic of younger people [61] and peak of SMS activity corresponding to the start of the school day, open lunch, and dismissal. Authors were able to establish conclusively that this cluster was mostly composed by students by ana-

lyzing the activity on the antenna that covers a school in the studied area and finding the same behavior.

In [62] authors propose a framework to characterize network-wide usage profiles and evaluate it using a CDR dataset of 5 million subscribers. The goal is to build categories of network usages from a raw CDR dataset, which, in this case, provides hourly number of calls per antenna. The framework creates a node from each hourly measurement and structures all nodes as a dendogram. A set of network usage profiles results from hierarchically merging similar nodes into clusters until a certain threshold defined by two stopping metrics, Beale and C-Index. The similarity between nodes relies on two proposed metrics, traffic volume similarity and traffic distribution similarity, which take into consideration absolute and relative call volumes on all antennas inside a geographical area. Authors show that more than being able to find a coherent set of voice call usage categories, their framework produces as by-product the identification of anomalous call usage profiles.

Literature presents efforts on the categorization of users by their network activities using other parameters, e. g., visited websites [63], and WiFi usage [13]. Additionally, [64, 65] present network-wide analysis for special days such as World Cup match, Easter, Christmas, and Carnival or special areas of a city [66]. Besides, with focus on urban planning, [67] characterizes the usage of urban area based on the network activities for typical and special days. Still towards better urban planning, [68] proposes a framework that merges city information from network activity, vehicle traffic, and events to provide a better understanding on how cities function and to develop more efficient urban policies.

Most studies in the literature on the analysis of network utilization are based on the analysis of calling patterns (i. e., generated only when a voice call or a short message service occurs) usually described in CDRs with no regards to the actual data traffic generated by smartphone applications (e. g., email checks, synchronization, etc). Such analysis may provide an idea on the activity of mobile network customers but do not describe realistic data traffic demand patterns.

2.4 CONCLUSIONS

It is undeniable the importance of datasets in the network and mobility analysis, they enable an atemporal data investigation of events that otherwise would be very hard to observe and study in a real time fashion. The main drawback is the lack of publicly available rich datasets in terms of fine-grained and network traffic information. As a consequence, this thesis tackles mobility and network traffic analysis using different datasets, which will be presented in the following

chapters. Our analysis from both mobility and network traffic differs from previously mentioned works in several aspects, we:

- focus on trajectory analysis and how people interact with the urban scenario
- provide large-scale fine-grained mobility study focusing routinary behavior
- analyze several cities at once, avoiding bias of a specific city context
- extensively analyze real cellular network data traffic and its cyclicity

With that in mind, we envision opportunities for better network management, and planning. Therefore, in Chapter 4 we propose a framework to automatically classify subscribers into a set of profiles based on their network usage, and a synthetic traffic generator that mimics the original network demands' behavior. Another opportunity that we explore is the deployment of a supportive wireless network on metropolitan areas, which we evaluate as study case in Chapter 5.

3 CONTEXT ANALYSIS

"We are what we repeatedly do."

— Aristotle

3.1 INTRODUCTION

On the pervasive era people are connected receiving and sharing information about themselves and their surrounding scenario. People actively or passively share details about their context and the content (as investigated in the next chapter), which are important sources of insights in order to understand human routine.

People are routinary semi-rational entities, they have regular circles of actions guided by their decisions but unexpected situations may interfere on their directions [69]. A person may change their itinerary due to a traffic jam, problems on the public transportation, etc. When choosing an itinerary, people tend to use the shortest-path to reach their destination, also known as *"desire line"*. The desire line is the shortest line between origin and destination, and expresses the way a person would like to go, if such a way were available [70]. Furthermore, the people's itinerary is characterized by its *confinement*, i. e., despite of choosing the shortest-paths, people will roam close by their main physical address [39].

Literature studies human mobility predominantly from coarse-grained datasets, from which its sparseness and coarseness plays a negative role on the understanding of some specific aspects such as the usage of shortest-path. Our main contribution in this chapter is to present an *extensive human mobility analysis from several fine-grained and one additional large-scale coarse-grained dataset*. Our datasets represent human mobility from 8 cities in 3 different continents around the world, namely London, Moscow, New York, Paris, San Francisco, Tokyo, Mexico, and Beijing. For each of them, we first model urban scenario with GPS- or CDR-based trajectories and points of interest (Section 3.2.1). Our points of interest represent real venues. We have collected information regarding more than 1.4 million unique venues distributed among the studied cities. Our human mobility evaluation comprises visit, temporal and spatiotemporal aspects (Section 3.3). From our analysis we show that human mobility presents three main characteristics: *tendency to use shortest-path, confinement and a strong repetitive behavior relative to few locations*.

3.2 RATIONALE

We analyze several different datasets aiming to find and measure a set of consistent characteristics present in the routine of human mobility. In order to avoid bias on the mobility aspects, the datasets come from different sources, cities and periods of time. Table 1 describes the characteristics of each of the datasets. The mobility datasets come from OpenStreetMap[2], GeoLife[24] and a Telco operator in Mexico. Therefore, they have different characteristics on how to represent user mobility. Before presenting our mobility analyzes, we provide in this section the insights considered in our datasets and a discussion about our system model. Due to space restrictions, graphics show results for Beijing, otherwise stated. However, we will highlight results from different cities throughout the discussion.

3.2.1 *System model*

For each of our datasets, we build a system model that represents a fairly real urban scenario composed by its respective users and their trajectories. Besides, we use data describing more than 1.4 million real points of interest spread in the cities we consider. Next sections describe the urban scenarios created with OpenStreetMap, GeoLife and a Telco dataset.

3.2.1.1 *Urban scenario*

GEOLIFE: We use the latest version of GeoLife dataset [24]. GeoLife is considered to be unique in the literature. This is due to the fact that it provides a rich view of people mobility using 11 different transportation modes in an urban area for a long period of time. It provides geolocalized and timestamped points from 182 people during a 4 year span, from 2007 to 2011, mostly in Beijing. For each person, the dataset provides a set of geolocalized points ascendingly sorted by timestamp, i. e., a fine-grained GPS trajectory. All components are based on geolocalized information, i. e., latitude and longitude coordinates within a 2004 km² central area in Beijing. Moreover, to better understand specific behaviors inherent from different periods of the day, every day is divided into four periods of time with 6 hours, from 00:00 to 05:59, from 06:00 to 11:59, from 12:00 to 17:59, and from 18:00 to 23:59. Those periods were chosen because they represent important parts of the day, late night, morning, afternoon, and evening, respectively.

Due to the routine behavior of people and the large time scale of the GeoLife dataset, it suffices to study a subset of the whole

2 http://www.openstreetmap.org

dataset in order to capture the daily behavior of subscribers. Since our work bases its premises on routinary behavior present on the mobility of people, a time subset of GeoLife data is already enough to show how we can explore routine to provide a better user experience. Therefore, we select the data of the two most active months in terms of number of users and trajectories. This subset spans from 1st November to 31st December of 2008 and contains 39 users and 2203 trajectories. The following results use this subset of data, unless stated otherwise.

OPENSTREETMAP: We have collected trajectories using the official OpenStreetMap API[3]. OpenStreetMap is a collaborative project with more than 1.9 million registered users. It has a feature in which users can upload their geolocalized trajectories in order to improve the mapping. We analyze about 14,200 public trajectories uploaded to OpenStreetMap from 6 cities, London, Moscow, New York, Paris, San Francisco and Tokyo. As in GeoLife, each user's fine-grained GPS trajectory is a set of geolocalized points ascendingly sorted by timestamp. Besides, similarly to GeoLife, the days were divided in periods of 6 hours each.

TELCO: Consists of a CDR dataset with about 6.8 million subscribers collected in a large urban area of Mexico city. It contains the geographic position of the antenna being used and the instant of time when the call was performed for each subscriber from July to October, 2013. As usual to CDRs, in general a few number of geographic points are present per user in each day due to the time sparsity of the calls. Moreover, due to the routinary behavior, people tend to make calls on the repeated areas, i. e., antennas. Thus, we have created a 1 week dataset from the original 4-month dataset as following: each day of the week has all geographic positions from this respective day throughout the original dataset, .e.g, Monday on the 1-week dataset has the GPS positions of all Mondays in the 4-month dataset for each user. Consequently, this dataset better represents the mobility of the subscribers with more geolocalized points.

3.2.1.2 *Trajectories*

A trajectory represents how people move around and it is described as a set of points representing GPS coordinates periodically collected. Regardless the dataset, each trajectory point has latitude, longitude and timestamp, to indicate when the position was recorded.

GEOLIFE AND OPENSTREETMAP: People may move around building their trajectories using at most ten different transportation

3 http://wiki.openstreetmap.org/wiki/API

Table 1: Characteristics of the mobility datasets

City	*Users*	*Period*	*Days*	*Source*
London	167	7th Nov., 2006 to 14th Dec., 2014	1073	OpenStreetMap
Moscow	197	4th Sep., 2005 to 17th May, 2014	1628	
New York	41	14th Feb., 2008 to 30th Oct., 2014	120	
Paris	182	19th Aug., 2007 to 8th Jan., 2015	556	
San Francisco	62	18th Apr., 2008 to 16th Sep., 2013	214	
Tokyo	87	10th Dec., 2007 to 13th Sep., 2013	513	
Mexico City	6.8 M	1st Jul. to 31st Oct., 2013	123	Telco
Beijing	182	12th Apr., 2007 to 27th Oct., 2011	1603	GeoLife

modes such as taxi, bike, run, bus, walk, train, subway, car, boat, and motorcycle. Therefore, to capture this urban behavior, we have divided each trajectory into *legs*. A leg is a subset of points from a trajectory with an unique transportation mode. For example, a trajectory composed by two legs, "car" and "walk" may represent a situation in which a person went by car to his work, parked the car and went walking until his office.

On the point of view of the GeoLife experiment, the transportation mode is a label set by the users being tracked. However, not every trajectory from the dataset was originally labeled with a transportation mode. That is due to the fact that labeling was not mandatory for people participating on the GeoLife experiment. Similarly, OpenStreetMap trajectories are not labeled at all. To overcome this limitation, we have created a model that labels legs by their speed compared against known average speeds for transportation modes.

In particular, we calculate the average speed of a leg. For Beijing, the result is matched against known speed ranges shown in Table 2. Lets consider we have an unlabeled leg l traveled with average speed of 5 m/s. It is possible to see that l falls on a range in which three transportation modes present over-

Table 2: Average speed ranges by transportation mode

Transportation mode	Average speed range (m/s)
Walk	$\leqslant 1.5$
Bus	> 1.5 and $\leqslant 4$ [71]
Bike (99.7%) or Run (0.3%)	> 4 and $\leqslant 4.4$ [72]
Taxi (39%), Motorcycle (0.1%), or Car (60%)	> 4.4 and $\leqslant 11.5$ [73]
Subway	> 11.5 and $\leqslant 28$ [74]
Train	> 28 and $\leqslant 250$
Airplane	> 250

lapping average speed ranges, "taxi", "motorcycle", and "car". In order to keep the proportion of legs that were originally labeled by the users in the experiment, we have calculated the percentage of legs (shown between parentheses) on each overlapping range. Therefore, in this example, l will be labeled either as "taxi", "motorcycle", or "car" with 39%, 0.1% and 60% of chance, respectively. For cities on OpenStreetMap, we apply similar methodology, but due to the lack of a sample with labeled trajectories from the original dataset, the probabilities are equally divided, e. g., "taxi", "motorcycle", or "car" have 33.3% of chance.

TELCO: As usual to mobility traces based on CDRs, we consider that the positions of the antennas whose subscriber is connected to during the day represent the trajectory points. For each subscriber the points are ascendingly sorted by time of the day. In our dataset, about 70% of the antennas are inside Mexico City urban area and the median pairwise distance between sequential trajectory points per subscriber is 1.5 km. It is much more coarse-grained than the 16 meters from GeoLife and from 7 to 18 meters in the cities from OpenStreetMap. Therefore, differently from the GeoLife and OpenStreetMap, no transportation mode is inferred from the trajectory points in this dataset.

Figure 3.2(a) shows the total number of users and trajectories they have performed in Beijing. As expected, *the number of users and the number of trajectories are highly correlated*. It is remarkable the similarity on the shape of the curves for both parameters. Indeed, Pearson's correlation between number of users and number of trajectories is 92%. Similarly, this correlation is 72% in Moscow and 70% in London.

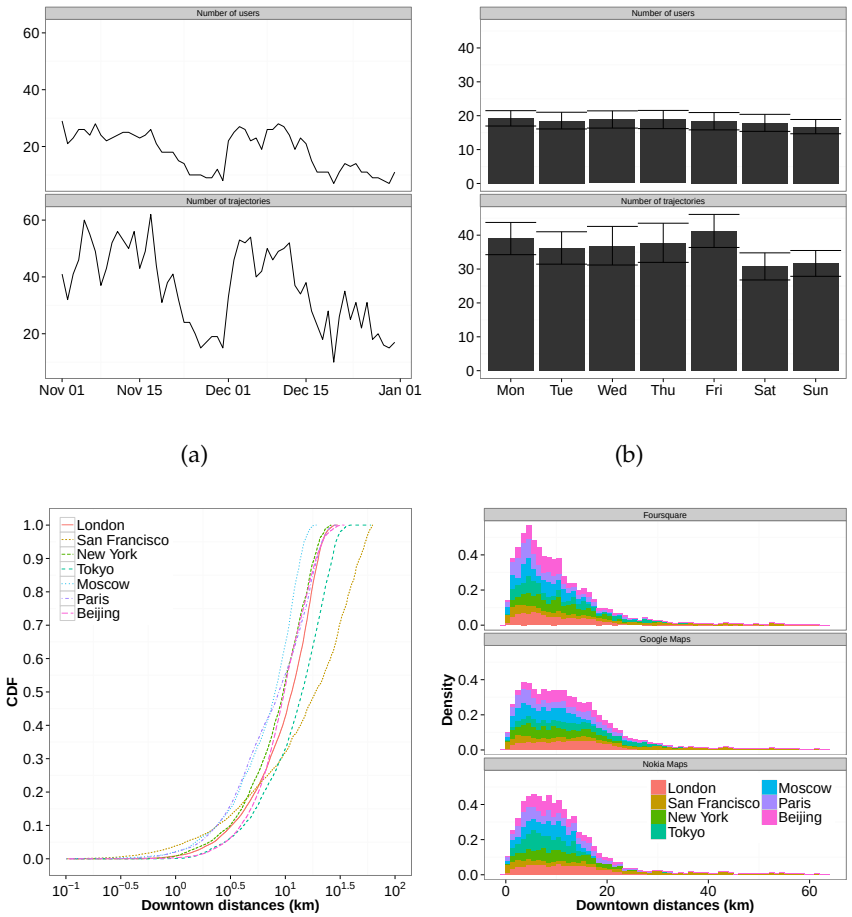


Figure 2: (Better seen in colors) (a) Number of users and trajectories per day in Beijing. (b) Number of trajectories per user grouped per day of the week in Beijing. (c) CDF of the distance from points of interest to the downtown. (d) Distance from points of interest to the downtown grouped by source.

In Figure 3.2(b), we present the average number of users and trajectories on each of the week days in Beijing. The day-wise difference for the number of users slightly decreases as the week progresses. On average, the highest difference is 12% more users on Monday than on Sunday. As expected, *week days present more people than on weekend days.* Indeed, for the week, the average difference is 7%. For the number of trajectories, the behavior is generally decreasing, except for Friday, which has the higher average number of trajectories when compared with all days of the week. That is probable due to people that go to their usual daily activities e. g., places related to work or study and added more trajectories related to leisure during the night such as bars or restaurants. On average, the day with the highest number of trajectories, which is Friday, presents 25% more trajectories than the one with the least number of trajectories, which is Saturday. All cities

presented similar results, for instance, London, San Francisco, Tokyo, and Moscow have 29%, 19% 35% 46% more trajectories on Friday than on Saturday. As expected, *on average, the number of trajectories is higher during the weekdays than during the weekend.* The difference is, on average, 2%, 12%, 23%, 9%, and 19%, respectively for London, San Francisco, Moscow, Paris, and Beijing.

3.2.1.3 *People and Points of Interest*

People move, build their trajectories, and carry mobile devices capable of Wi-Fi communication and able to receive GPS information. While walking by, people may "interact" with points of interest (PI), e. g., bar, bus station, supermarket, etc. Those points of interest describe more than mere locations in the map but they reflect a social aspect e. g., students are frequently going to meet their colleagues in a coffeehouse close to the university they attend to. Indeed, this represents a routinary behavior that involves not only people but also their interaction with points of interest in a city. In our scenario, points of interest are geolocalized physical venues generally present in most of major cities, e. g., bank, supermarket, cafe, school, stadium, train station, university, etc.

In order to study the deployment with real points of interest in Beijing, we have collected information from different databases of places (e. g., Google Places[4]). Such databases are growing and are the most accurate source of public information about points of interest. To avoid to be biased by the characteristics of one unique database, we have collected data from multiple sources namely Google Places, Nokia Maps[5], and Foursquare[6], counting more than 1.4 million real unique points of interest with their respective IDs, latitudes and longitudes. While Google Places and Nokia Maps databases provide information about points of interest collected from city hall's records, by the respective owner of the venues, etc, Foursquare provides only information from places where its users *checked in*[7], generally places related with leisure and social relationships. For each set of places collected from a source, repeated ones were removed by keeping an unique occurrence of each ID. Table 3 describes the characteristics of the sets of collected points of interest per source in 98 categories, e. g., market, library, school, etc. It is important to enhance that due to considerably higher spatial distance between trajectory points from the Telco dataset when compared to GeoLife and OpenStreetMap, we have chosen to not include points of interest for Mexico City.

Figure 3.2(c) shows the distribution of points of interest by their distance to each of the cities' downtown. For instance, we have con-

4 https://developers.google.com/maps/documentation/places
5 https://developer.here.com
6 https://developer.foursquare.com
7 https://support.foursquare.com/hc/en-us/articles/201065340-Check-ins

sidered the downtown as being centered at the Big Ben for London, Market Street for San Francisco, Central Park for New York, Imperial Palace for Tokyo, the Red Square for Moscow, Île de la Cité for Paris and Forbidden city for Beijing. Regardless the city, there is a concentration of points of interest closer to the downtown. For instance, 52%, 65%, 54%, and 56% of the points of interest are located at most 10 km from the downtown for Beijing, Moscow, New York, and Paris. Figure 3.2(d) shows histograms of the distances from the points of interest to city downtown grouped by source, Foursquare, Nokia Maps, and Google Maps. This result shows that, regardless the source, *points of interest are more concentrated closer to downtown*. Indeed, considering all cities, the highest concentration for Foursquare, Nokia Maps, and Google Maps falls into (2, 4.17], (6.2, 8.3], and (2, 4.18] km, respectively. Moreover, the median distance from points of interest of Foursquare, Nokia Maps, and Google Maps to the downtown is 8.8, 9.6, and 11.4 km, respectively. Since Foursquare venues are mostly related to leisure, they tend to be, on average, closer to downtown than the ones from Nokia Maps and Google Maps, whose points of interest are distributed in a wider range of niches.

Table 3: Area and number of points of interest per city

City	*Area (km²)*	*Points of Interest*		
		Google Maps	*Nokia Maps*	*Foursquare*
London	1747	227757	56434	44469
Moscow	645	65712	34795	8659
New York	836	88608	61167	33690
Paris	1725	193237	41476	18767
San Francisco	2433	131470	36677	37901
Tokyo	2288	155696	7954	5415
Mexico City	5515	-	-	-
Beijing	2004	77919	119346	5059

3.3 MOBILITY DYNAMICS

So far we have shown isolated characteristics of the dataset such as number of users, trajectories and how PIs are arranged in the various cities. On the other hand, this section presents several analysis of the relation between people's mobility and urban scenario considering time and space, e. g., time spent on shopping areas during week and weekends.

In order to understand the routine aspects on the mobility of people and its interaction with the urban environment, some of the following analysis consider data summarized by period of the day or day of the week. For instance, the curve labeled *Monday* on Figure 3.3(a) shows the cumulative distribution for the average number of unique users per PI for all Mondays. This also applies for the periods of the day, whose data represents all occurrences of the respective measurement for each of the periods in all days of the week. Our description of the results per day of week considers that the week *progresses* (or *passes by*) from Monday to Sunday, i. e., it follows the ISO 8601 [75]. Additionally, we refer *weekend* as Saturday and Sunday together.

3.3.1 *Visit behavior*

This section assess how people interact with the urban scenario. The *interaction* is a broad term and our analysis explores several aspects which can reveal our daily routines. Since Telco dataset does not contain PIs, the results in this section do not take Mexico City into consideration.

Figure 3.3(a) shows the CDFs of the number of unique users per PI, i. e., number of users that visited a PI in Beijing. Note that we count a visit to a PIs if a person enters on its interaction range (Sec. 5.3.3.1). Due to the large number of the PIs in the city, most of the PIs are rarely visited by the users on a single day. Indeed, 78% of the PIs are visited by only one user per day. This holds for each of the days of the week, 92%, 92%, 91%, 92%, 91%, 94%, 96% of the PIs are visited by up to two users on Monday, Tuesday, Wednesday, Thursday, Friday, Saturday and Sunday, respectively. All the other cities present similar results. Moreover, PIs receive 9% less users on weekends when compared to weekdays. For Moscow, 93% of the points of interest are visited once. This percentage is even higher for the other cities. Although GeoLife has less users than OpenStreetMap in Moscow, the former tends to have a higher number of users per day, which increases the number of unique visits per PI.

The number of visits is an interesting parameter that may undercover, for example, how frequent people tend to visit a certain PI, but it does not show an important characteristic of the mobility: How long people stay in the vicinity of a PI. Most of the PIs present low

coverage time during the day, 49%, 63%, 52%, 29%, 75%, 56%, and 62% of the PIs have users on their vicinity for up to 100 seconds and 80%, 87%, 87%, 50%, 92%, 88%, and 87% up to 1000 seconds in London, San Francisco, New York, Tokyo, Moscow, Paris, and Beijing, respectively. Figure 3.3(b) shows the coverage time per PI, i. e., the total amount of time that users have spent inside its interaction range per day of the week in Beijing. There is a clear increasing behavior of how long people stay in the vicinity of the PIs as the week passes by. That is probably due to the hurry of the metropolitan areas, in which people start the week in a higher pace than they end. People pass by venues as quick as possible but as the time goes towards the weekend they have more time to spend, e. g., looking at the showcases or visiting stores. On average, *people spent 63%, 87%, and 33% more time on the vicinity of places on Sunday than on Monday* for San Francisco, Beijing, and Moscow. Additionally, *PIs, on their vicinity, have people 12%, 20%, and 17% more time on weekend than on weekdays* for the same cities.

Figure 3.3(c) shows the CDFs of the total number of PIs seen, i. e., including repeated visits to the same PI, per user per periods of the day in Beijing. The earliest and the latest periods of the day, i. e., from 00:00 to 05:59 and from 18:00 to 23:59, present the least number of visited PIs. Briefly, the main reason for that is the shorter length of users' trajectories during those two periods compared with the other periods of the day. We further develop the idea of trajectory length in Sec. 5.3.2. On the other hand, *users visit the highest number of PIs from 12:00 to 17:59*, which is expected due to the daily activities. On average, users on the period from 12:00 to 17:59 visit 82% and 79% more PIs than from 00:00 to 05:59 and from 18:00 to 23:59, respectively. In Tokyo, for the same periods this difference is 57% and 60%, respectively. Additionally, if we consider the majority of the users, for example, up to up 75% of them in Beijing, 94% more PIs are visited per user from 12:00 to 17:59 than from 00:00 to 05:59, which is the period with least number of PI visits per user. Still comparing in Beijing, for the same percentage of users, 52% more PI are visited per user from 12:00 to 17:59 than from 18:00 to 23:59. Regarding the period from 06:00 to 11:59, which is also almost as active as from 12:00 to 17:59, the former has 11% less visits per PI than the latter.

New York presents similar results, for up to 75% of users, the most active period of the day, which is from 12:00 to 17:59, has 27%, 72%, and 21% more visits to PIs than from 00:00 to 05:59, from 06:00 to 11:59, and from 18:00 to 23:59, respectively. It is important to note that this behavior slightly varies in certain cities, while it presents extremely similar results in others. Aside from New York, Tokyo presents very similar results. On the other hand, cities such as London and San Francisco still had high number of visits during daylight periods, but similar number of PI visits during the night. In such cases it is difficult to indicate a single reason. For example, it might be due to

the city context, i. e.people are as active during the day as during the night or to the dataset characteristics, which contain a more balanced number of users during the day and night.

Figure 3.3(d) shows the CDFs of the number of PIs seen per user per days of the week in Beijing. *Similar to spending more time on the vicinity of PIs during the weekends than on weekdays (Figure 3.3(b)), on our data, people tend to pass by more PIs on weekends than on weekdays.* On average, people visit 25%, 8%, 19%, 54%, and 72% more PIs on weekends than on weekdays in Beijing, London, San Francisco, New York, and Tokyo, respectively. Moreover, there is a growth on the number of PIs from Monday to Friday. Indeed, *on average 36% more PIs are visited on Friday than on Monday* considering all cities. Cumulative results show similar tendency, 30%, 21%, 40%, 65%, 25% more visited PIs on weekends than on weekdays for up to 75% of the users in London, San Francisco, New York, Tokyo, and Beijing, respectively.

We further investigate the interaction between people and urban scenario by segmenting the city using *cells*. In our context, cells are square-shaped regions of $50m^2$ organized in a grid fashion on the city terrain. Figure 3.4(a) shows the *hexagonal bin plot* [76] of cell distance to downtown and number of PIs inside the cell in Beijing. The intensity of a bin represents the frequency of cells that contain a number of PIs laying within the bin. There is a *densification, i. e., higher concentration of PIs closer to downtown, that decreases with the increase of the distance.* This is a common aspect of metropolitan areas, there is a strong negative correlation between distance to downtown and number of venues, -95%, -55%, -64%, -57%, -90%, -97%, -93% for London, San Francisco, New York, Tokyo, Moscow, Paris, and Beijing. Although San Francisco and Tokyo are big metropolitan, their concentration on the surface tend to be more truncated due the limitations of the bays present on both of them. That is the probably cause of their lower correlation compared to the other evaluated cities. There is also a high frequency of cells containing a low number of PIs irrespective of the distance to downtown. Bigger venues may explain this, e. g., a city hall could occupy the whole space of a single cell.

Figure 3.4(b) shows the total number of visited cells, i. e., including repeated visits to the same cell, per user per day of the week in Beijing. The tendency of the CDF curves is similar to the Figure 3.3(c), but shifted to the left due to an expected lesser number of cells than PIs. Additionally, *on average the number of visited cells grows from Monday to Sunday with a peak on Friday.* For instance, people visited 65% more cells on Friday than on Monday and 6% more cells on weekends than on weekdays. When considering all cities, on average, those percentages are 55% and 12%, respectively.

In order to better understand the predictability of people's mobility, we calculate the L rank [41] of the visited PIs and cells. The rank is calculated per user and it takes into consideration the number of

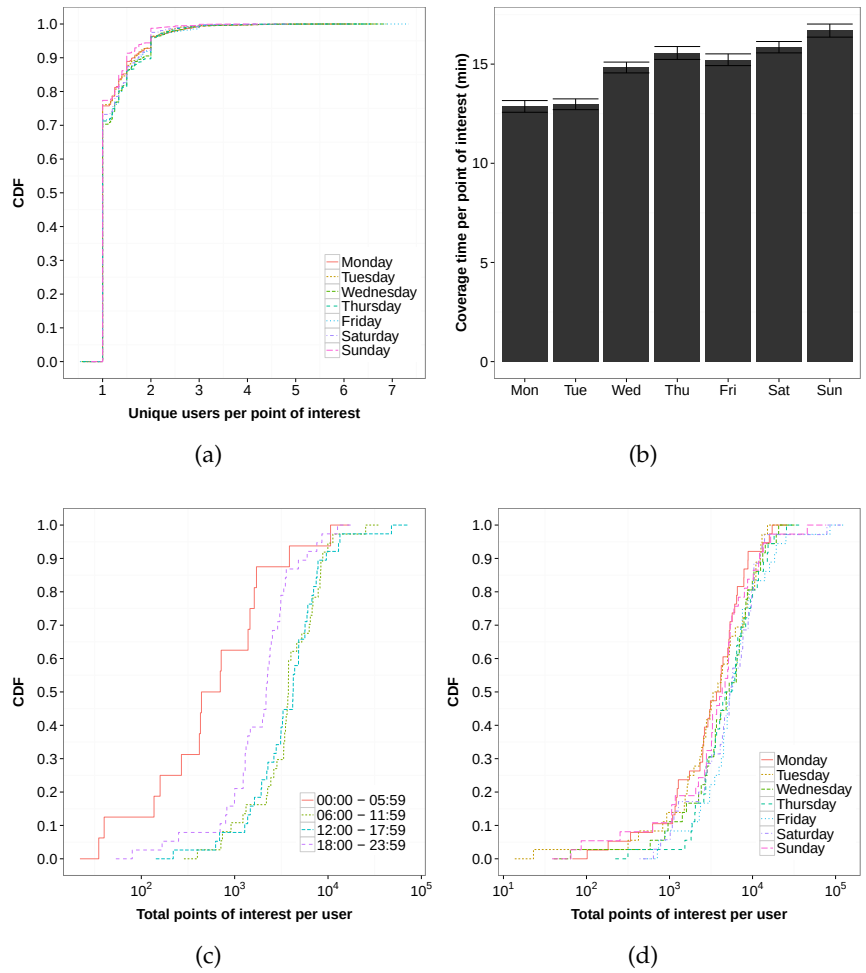


Figure 3: (Better seen in colors) (a) Number of users and (b) mean coverage time per PI and day of the week in Beijing. Number of visited PIs per user per (c) period of the day and (d) per day of the week in Beijing.

times he visits a cell/PIs, e. g., the most visited cell/PI by a user has rank $L = 1$. Figure 3.4(c) shows a Zipf plot of the visiting frequency for the cells and PIs ranked L in San Francisco and Beijing. The dashed straight line shows that the distribution can be approximated by $1/L$. Furthermore, *this plot shows that users concentrate most of their visits to few frequently visited cells and PIs, i. e., to a very restricted area.* For example in Beijing, 43% and 40% of the visits are made to 1% of the cells and PIs, respectively.

From the previous analysis, we see a strong repetitive tendency on the human mobility, i. e., repetitively visit the same areas. To quantitatively express how repetitively visited is a PI, a metric called *Repetitiveness* was conceived. The repetitiveness of a PI v_i is based on the

number of unique users (NUU) and total users (NTU) that visited it as following:

$$re(v_i) = \frac{NTU_{v_i} - NUU_{v_i}}{NTU_{v_i}} * 100 \quad (1)$$

Figure 3.5(a) shows the CDF of the average repetitiveness for each of the days the week in Beijing. This result shows that the majority of the PIs present low repetitiveness and a minority has high repetitiveness. Indeed, 23%, 30%, 30%, 13%, 23%, 30%, and 23% of the PIs in London, San Francisco, New York, Tokyo, Moscow, Paris, and Beijing present repetitiveness up to 50%. For the same cities, 47%, 58%, 56%, 26%, 56%, 55%, 55% of the PIs have up to 85% of repetitiveness. Besides, for all cities $\approx$ 1% of the PIs are highly repetitively visited, presenting more than 98% of repetitiveness. As the system model for Mexico city does not contain PIs, we have calculate repetitiveness using the cellular network's antennas. Figure 3.5(b) shows the CDF of the average repetitiveness for each of the days the week in Mexico considering cellular antennas instead of PIs. The results are similar to counterpart results for PIs, majority of antennas are barely used, while a very small amount of them is used several times by the same subscriber. 6% of the antennas present up to 50% of repetitiveness, 90% have up to 85% of repetitiveness, and less than 1% of antennas present more than 98% of repetitiveness. For all cities, the average repetitiveness does not present significant differences between weekdays and weekends, 4.5%.

3.3.2 *Displacement behavior*

In order to evaluate how much space, i. e., the physical extent, users cover on their mobility and how they travel on the space, we evaluate the *radius of gyration* r_g [41], maximum displacement, and the desire lines of users' trajectories. Radius of gyration is the linear size occupied by each user's trajectory up to time t and it is formally defined as:

$$r_g^u(t) = \sqrt{\frac{1}{n}\sum_{p=1}^{n}(\overrightarrow{r}_p^u - \overrightarrow{r}_{cm}^u)^2} \quad (2)$$

where $\overrightarrow{r}_p^u$ represents all the trajectory points $p = 1, \ldots, n$ of the user u and $\overrightarrow{r}_{cm}^u = \frac{1}{n}\sum_{p=1}^{n}\overrightarrow{r}_p^u$ is the center of mass of the trajectory. $r_g^u(t)$ captures thus how broadly the users travel as opposed to the actual distance traveled. In our results, the unit of the radius of gyration is meters. During t all the trajectories of each user u are considered in the calculation of r_g, which is expected to grow with the growth

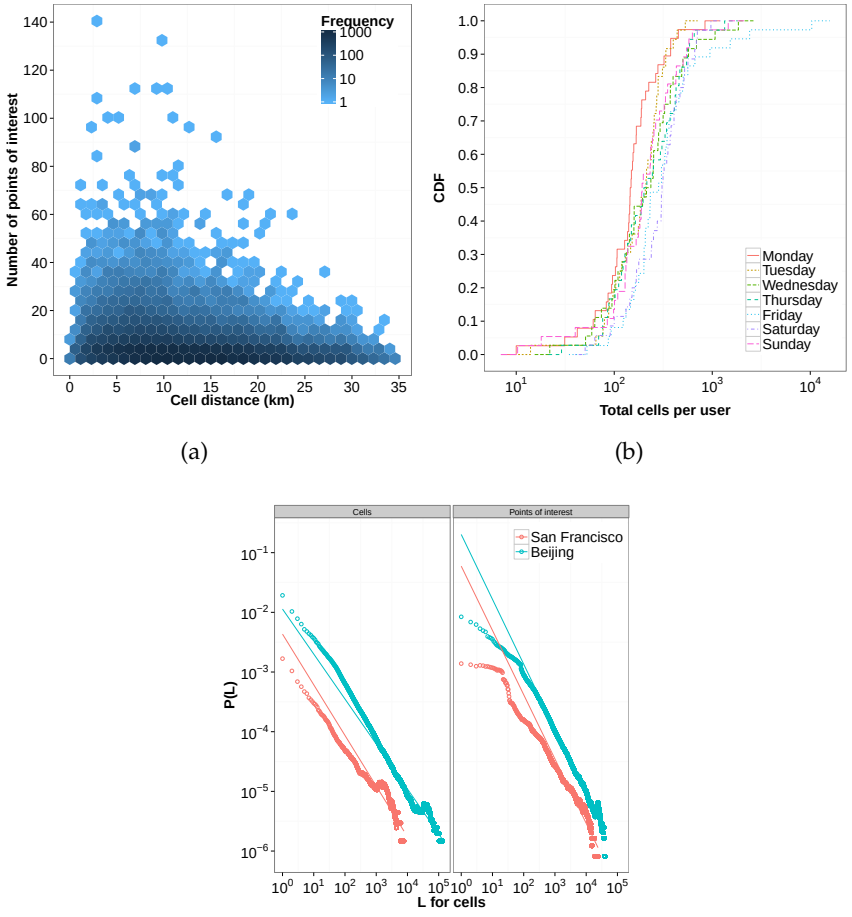
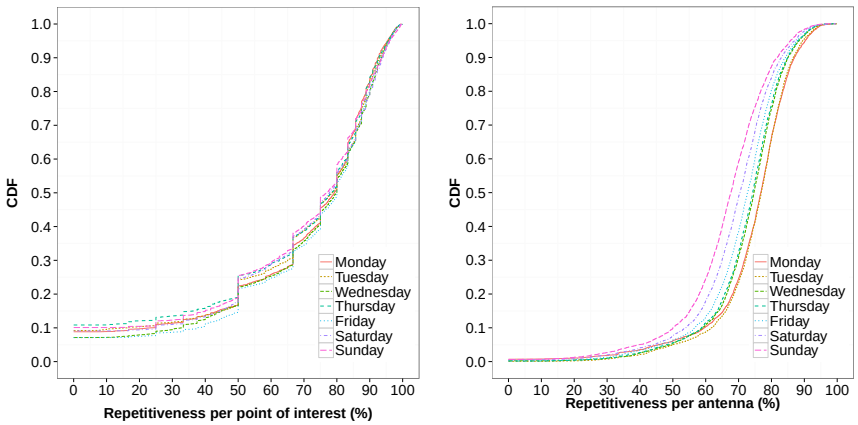


Figure 4: (Better seen in colors) (a) Bin plot of the number of PIs per cell distance in Beijing. (b) Total number of cells visited per user per day of the week in Beijing. (c) L rank for cells in San Francisco and Beijing.

of t. The following results analyse r_g for different durations of time, periods of the days, whole days and the 2 months of the dataset.

Figure 3.6(a) shows the CDFs of the radius of gyration per period of the day on weekdays and weekends. *On weekdays, the earliest period of the day presents the smallest radius of gyration.* That is coherent with human routines, from 06:00 to 23:59 we perform more activities and are more susceptible to displacement that covers a wider area. Contrarily, *from 00:00 to 05:59 people are more stationary performing at most short trajectories and likely at home, sleeping, etc.* Median radius of gyration per user in the period from 00:00 to 05:59, is 92% shorter than the radius of gyration from 06:00 to 11:59 in Beijing. Taking into consideration all cities, except Mexico, the radius of gyration from 00:00 to 05:59, is 53% shorter than the radius of gyration from 06:00 to 11:59. For instance, the average radius of gyration is 759 meters in



(a) (b)

Figure 5: (Better seen in colors) Repetitiveness of PIs per days of the week in (a) Beijing and (b) Mexico.

from 00:00 to 05:59 and about 10 km from 06:00 to 11:59 in Beijing. Due to the sparsity of the Telco dataset, the radius of gyration tends to be larger on Mexico than on the other cities. Therefore, we have made a separate analysis for Mexico dataset on displacement aspects. 3.7(a) shows the radius of gyration for subscribers per period of the day on weekdays and weekends in Mexico. Radius of gyration from 00:00 to 05:59, is 69% shorter than the radius of gyration from 06:00 to 11:59. For instance, the radius of gyration is 13.9 km from 00:00 to 05:59 and 46.4 km from 06:00 to 11:59.

Differently, on weekends the radius of gyration from 00:00 to 05:59 grows 49% in Mexico, 46% in New York, 6% in Tokyo, and 35% in Beijing when compared to weekday. That is due to the nightlife activities which increase the late night mobility of the users. As *a probable consequence of the higher mobile behavior on weekends' late nights, there is a reduction on the average radius of gyration from 06:00 to 11:59 on weekends when compared to same period on weekdays.* For instance, it is 41, 2.5, 1.1, and 4.6 km in Mexico, New York, Tokyo, and Beijing, which it is 12%, 20%, 49% and 53% less than on the same period on weekdays, respectively. That is likely due to the people waking up later on weekends than on weekdays.

Figure 3.6(b) depicts the CDFs of the radius of gyration per day of the week in Beijing. On average, *users tend to journey over a larger area as the week passes by from Monday to Saturday, with the exception of Sunday which has* r_g *comparable to Monday.* For instance in Beijing, the radius of gyration on Monday, Wednesday and Friday is 1.6, 1.8 and 2.3 km, respectively. Considering all cities, except Mexico, it is 1.7, 1.8, and 2.2 km, for the same days, respectively. The radius of gyration has the highest values on Friday and Saturday, the latter 2.9 km in Beijing and 2.8 km average considering all cities except Mexico.

Due to the peak on Saturday, average radius of gyration is higher on weekends than on weekdays, 1.9 and 2.5 km, respectively in Beijing and 1.9 and 2.2 km for all cities. Figure 3.7(b) shows the same results for Mexico. The behavior present on the GeoLife and OpenStreetMap is also present on the Telco dataset, i. e., peak on Friday and Saturday, 69 and 71 km, respectively. Moreover, the average radius of gyration grows as the week passes by.

Figure 3.6(c) shows the CDF of the final radius of gyration per user in Beijing. The steady and constant increase on the CDF curve shows that *users are almost equally distributed by their radius of gyration*. To further analyze the radius of gyration, we have grouped users by their final radius of gyration into four groups: $r_g \leqslant 10^4$, $10^4 < r_g \leqslant 10^5$, $10^5 < r_g \leqslant 10^6$, and $r_g > 10^6$ meters. Figure 3.6(d) shows the average radius of gyration for each of the groups up to the hour on the x-axis. The confidence intervals are shown as shadows around the curves. The saturation on the curves shows an upper bound for the movement area on each of the groups. An interesting aspect is how fast each of the groups reach (or approaches) their saturation values. For instance, at the end of the first day, 69%, 17%, $\approx$ 1,% and $\approx$ 1% of the final r_g has been reached in the groups $r_g \leqslant 10^4$, $10^4 < r_g \leqslant 10^5$, $10^5 < r_g \leqslant 10^6$, and $r_g > 10^6$, respectively. On one week, users on the same groups have reached 88%, 40%, 35%, and 12% of their final r_g. It means that *users which mobility is more confined tend to reach the upper boundary of their movement proportionally faster than the ones who journey over larger areas.*

The concept of *desire lines* states that people tend to choose the shortest-paths to arrive on their destinations. In order to verify that, we have compared the length of each traveled leg against the length of the corresponding shortest path considering the same initial and final points of the original leg. Dividing the length of the original leg by the length of the shortest path allows us measuring how longer the path made by a person is from the shortest path. We have used Google Directions API[8] to compute the shortest path. The API receives the coordinates of both initial and final points and a *travel mode*, i. e., transportation mode. Then, it returns the shortest path considering the restrictions imposed by the existing routes and obstacles in the city for a specific transportation mode. Note that, we only have considered transportation modes where people have decision control of their paths. This excludes for example, buses, boats or trains.

Legs traveled by *walk*, *run*, and *bike* had their lengths divided by their respective shortest paths computed while using the API in *walking* mode. Google Directions API indeed has a *bicycling* travel mode, but at the moment, it does not contain routes in Beijing. Therefore, we use *bicycling* mode when available in the evaluated cities. Legs trav-

8 https://developers.google.com/maps/documentation/directions

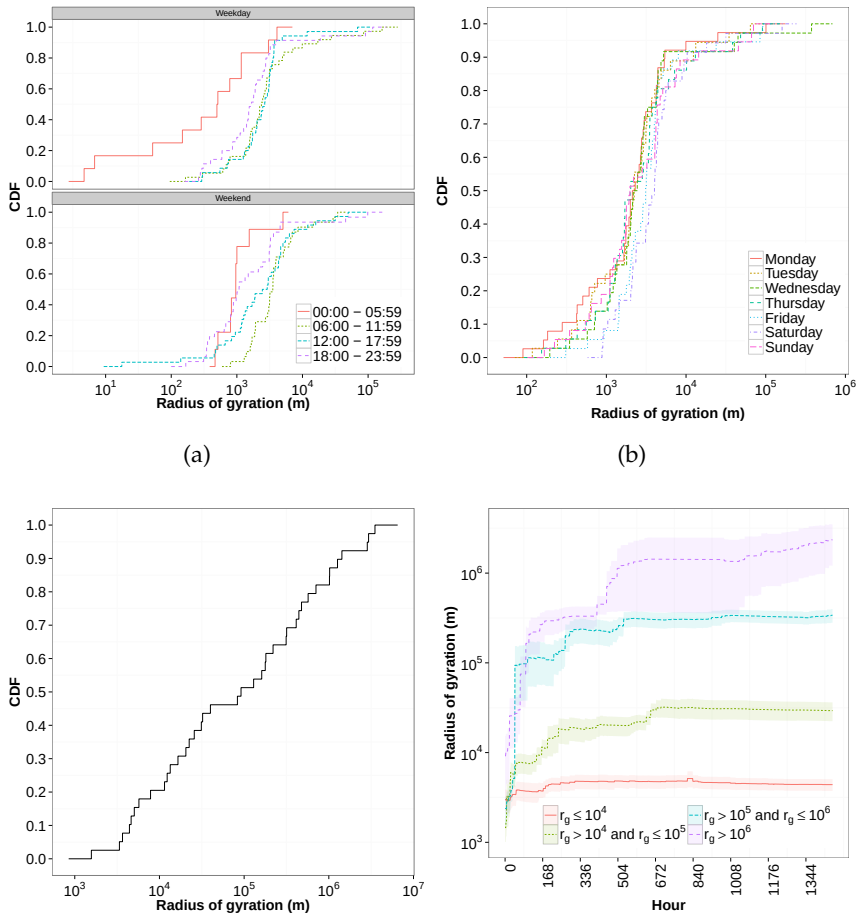
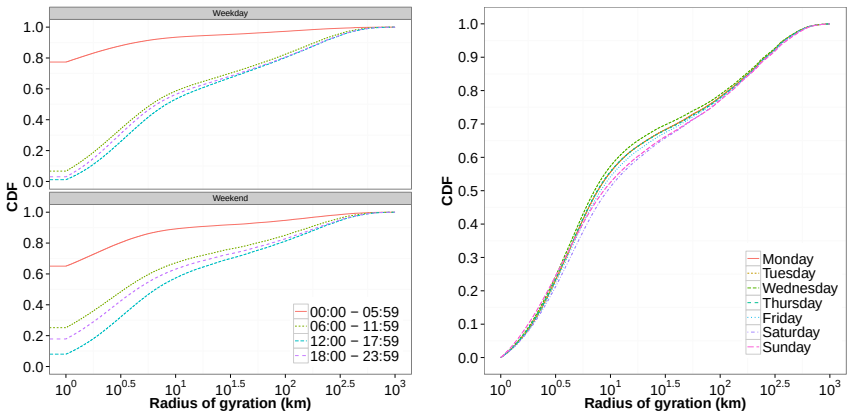


Figure 6: (Better seen in colors) Radius of gyration for users per (a) period of the day, (b) day of the week, (c) for all periods and days, and (d) for all periods and days grouped by final radius of gyration in Beijing.

eled by *taxi*, *car*, and *motorcycle* had their lengths divided by results of the API in the *driving* mode.

Figure 3.8(a) shows the CDF of the ratio between the original legs length and the shortest path, by transportation mode and period in Beijing. It shows that the periods from 00:00 to 05:59, from 06:00 to 11:59, from 12:00 to 17:59, and from 18:00 to 23:59 present, respectively, 36%, 62%, 52% and 74% of the legs measuring, at most, half longer than the shortest path. For all other cities, the average percentages for the same periods are 44%, 53%, 57%, and 73%, respectively. That results show that on late night people tend to walk around not directly going to their destination. Indeed, on late night people tend to go for bars, night clubs and are more susceptible to create routes that are way longer than the shortest ones. On the other hand, periods representing early morning and early night show high percentage of



(a) (b)

Figure 7: (Better seen in colors) Radius of gyration for subscribers per (a) period of the day and (b) day of the week in Mexico.

legs closest to the shortest one and describing how people go directly to their destinations, e. g., work, home, etc. The period containing the early afternoon hours present an intermediate percentage of legs close to the shortest path. Indeed, this period mixes people walking around careless about shortest paths (e.g, someone shopping, or looking for restaurants), and people more concerned about their being on time (e.g, people coming back from the lunchtime towards the work). Moreover, it is possible to see that the length ratio changes in function of the transportation mode. For instance, *walk* and *taxi* modes present the trajectories that are closer to the shortest path. That is probably due to the human capacity of being able to identify the trajectories, mainly when one have the knowledge of the neighborhood, which is the case of the participants of GeoLife experiment. Taxis tend to be equipped with GPS-enable devices and route planning software in order to find the addresses and the better (shorter in time and/or cost) routes. We conclude that *regardless of the transportation mode, people tend to be oriented by the shortest paths.*

We have also analyzed the length ratio grouped per weekdays and weekends and per transportation mode. For Beijing, the median length ratio is 1.1 and 1.3 on weekdays and weekends, respectively. Those values are consistent for all cities, on average 1.1 and 1.2, respectively. Additionally, on all cities, *bike*, *car*, *taxi*, and *walk* presented median length ratio of 1.05, 1.07, 1.12, and 1.39 for weekdays and 1.06, 1.2, 1.33, and 1.72 for weekends, respectively. It shows that *people on our datasets were presenting more routes closer to the shortest ones on weekdays than on weekends.* That is interesting because it measures a difference on people's behavior on weekdays and weekends.

People's mobility is generally *confined*. Even if people are not using the shortest routes, they are at least not going far from their home

location. To check how that premise occurs on our scenario, we have measured how confined the trajectories are by their maximum displacement. Maximum displacement is the distance between a trajectory's initial and farthest point (not necessarily the last point). Figure 3.8(b) presents the CDF of the maximum displacement for all trajectories grouped by period of the day in Beijing. It shows that *90% of the trajectories per period of the day have, at most, 10 km maximum displacement in Beijing. Considering all cities, except Mexico, this values is 75%.* Similar findings are present on the analysis of maximum displacement per day. For Mexico, the displacement is generally larger than 10 km due to the coarse-grained nature of the CDR dataset. Figure 3.8(c) presents the maximum displacement per period of the day for Mexico City. 57% of the trajectories per period of the day have, at most, 10 km. This value is lower than the other cities because of the sparsity of the dataset. For instance, 75% of the trajectories have, at most, 35 km. Additionally, there is a significant difference between the maximum displacement from 00:00 to 05:59 to the other periods, which is not observed nor on GeoLife neither on OpenStreetMap. This is due to the difference between a fine-grained mobility and sparse mobility. Generally, mobility is more frequent than calls from 00:00 to 05:59.

Figure 3.8(d) shows the CDF of the maximum displacement per user per day of the week in Beijing. For example, the median maximum displacement from Sunday to Thursday ranges from 4.4 to 5.2 km, and it is higher on Friday and Saturday, 7.2 and 7.9 km, respectively. This is a reoccurring behavior in all the cities. Median maximum displacement from Sunday to Thursday ranges from 6.3 to 6.9 km, and on Friday and Saturday, 7.1 and 7 km, respectively. This result shows that generally people do not move far away from their starting point, and presents high confinement. This result reinforces the findings for the radius of gyration. In fact, *there is a 96% correlation between users' maximum displacement and radius of gyration.*

3.3.3 *Spatiotemporal behavior*

People's mobility and visiting behavior may reflect some of their preferences and lifestyle. To better understand the visiting behavior of people, we have classified the categories of our PIs (refer to Sec. 3.2.1.3) in 9 classes. This classification aims to group together, on the same class, PIs whose categories carry similar meaning. For example, class *Education* (which we will call *Edu*) groups together all PIs with category "school", "university", and "library". Similarly, the remaining 95 categories were classified into more 8 classes. Table 4 describes the classes and some of the categories they contain. Refer to Section A.1 for the complete set of classes and their respective categories.

Figure 3.9(a) shows the amount of coverage time each of classes provided per period of the day in Beijing. Among all classes, *Food*,

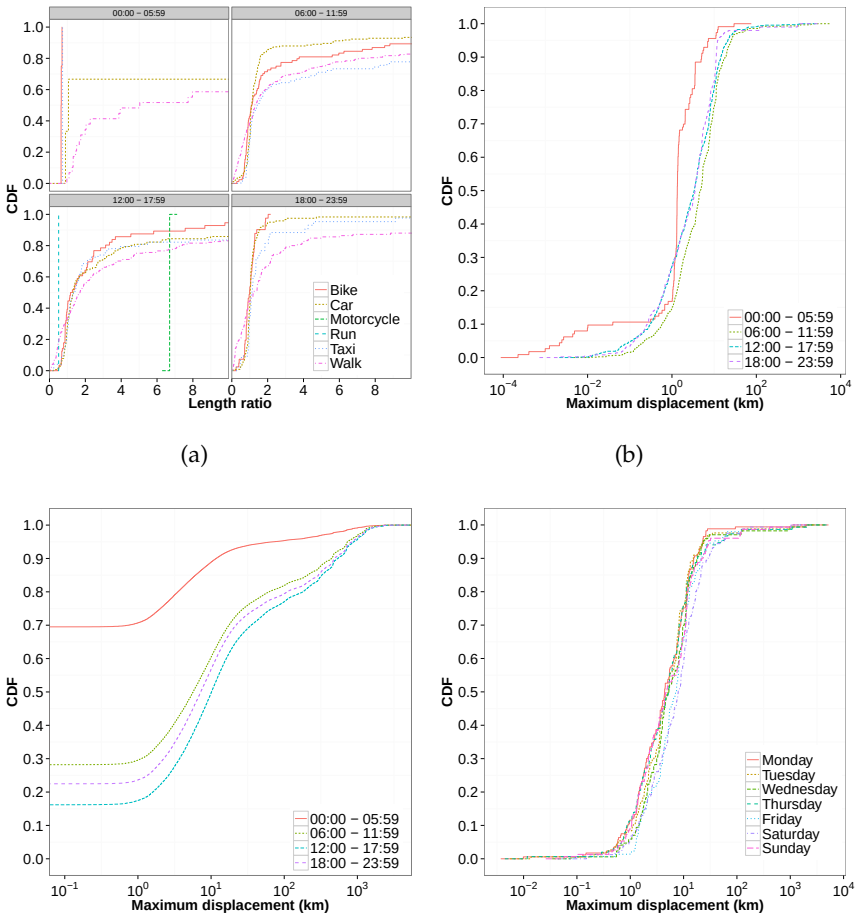


Figure 8: (Better seen in colors) (a) CDF of the length ratio per transportation mode grouped per period in Beijing. (b) Maximum displacement per period in Beijing. (c) Maximum displacement per period in Mexico. (d) Maximum displacement per day of the week in Beijing.

Shop, *Trvl*, and *Rel* present higher values on at least one period of the day. From 00:00 to 05:59, PIs in the *Food* class presents the highest coverage, 30 minutes on average. On London, San Francisco, and New York this period presents highest coverage for class *NL*, which is understandable due to the night-life related activities in those cities. From 06:00 to 11:59, *Shop*, *Food* and *Trvl* have the highest amount of coverage time, 115, 61, and 51 minutes, respectively. That is probably due to the shopping and breakfast-related PIs before using the transportation to work or study-related places. Similarly, from 12:00 to 17:59, the order of the classes with the highest coverage is the same: *Shop*, *Food* and *Trvl*, 211, 100, and 64 minutes, respectively. All other cities present mostly similar results from 12:00 to 17:59, with the inclusion of *Srvc* class being significant together with *Shop* and *Food*, they are among the top 4 classes that more provide coverage. This is

probably related to lunch and transportation back home. Differently, from 18:00 to 23:59, *Rel* class in Beijing shows the highest coverage time, 93 minutes, with large confidence interval. It is still unclear why the average coverage time for *Rel* is the highest in this period, but the large confidence interval is due to the few occurrences of this class. It might be the case that few users share a particular religious ceremony during night time. This period is not similar with any of the other cities we have evaluated and it is likely result from local circumstances. Aside from that, from 18:00 to 23:59 *Shop* and *NL* have 50 and 30 minutes of coverage time, respectively. For all other cities, those two classes are among the top 3 that most provided coverage in this period.

Figure 3.9(b) presents further investigation for the coverage time per points' of interest class. It shows the same data presented on Figure 3.9(a) grouped by weekdays and weekends instead of periods in Beijing. It is possible to see a significant difference between the coverage time on weekdays and weekends for *Shop* class. For instance, on average, it is 87 minutes on weekdays and 323 on weekends. On weekends, people normally have more time to spend on shopping areas than on weekdays and that is the probable cause of this difference. However, this aspect depends on the opening hours during weekends, e. g., *Shop* class in Paris class has, on average, 404 minutes of coverage on weekdays and 137 during weekends. Paris during weekends has a very limited number of venues opened compared to weekdays. Similarly, *Food* class has higher average coverage time on weekends than on weekdays in Beijing. *On metropolitan areas, shopping malls tend to concentrate food- and shopping-related venues, which it is likely the reason behind those two classes having high coverage time both on weekdays and weekends.* Furthermore, *NL* has 22%, 40%, and 8% higher coverage time on weekends than on weekdays for New York, Tokyo and Beijing, respectively, which is expected due to people having more time to spend on night life-related venues than on weekdays.

To further comprehend how people together explore the city on a spatiotemporal fashion, we use Moran's I spatial autocorrelation index [77] on snapshots of time. Spatial autocorrelation evaluates the correlation of variables among nearby locations in space and Moran's I can be defined as:

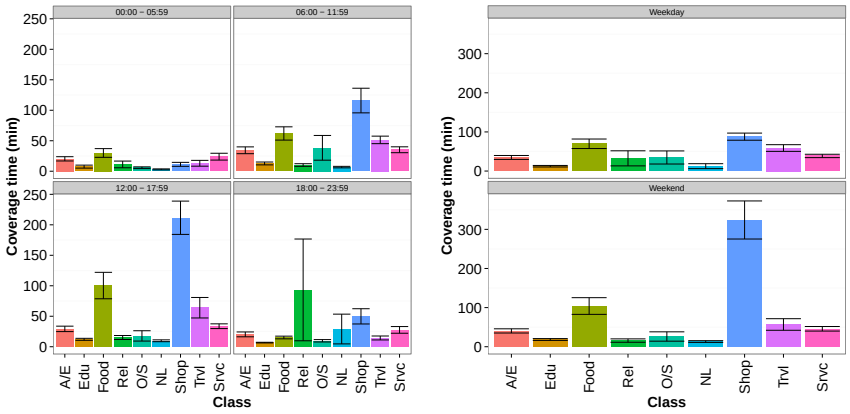
$$I = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(X_i - \bar{X})(X_j - \bar{X})}{\sum_i (X_i - \bar{X})^2}, \tag{3}$$

where N is the number of locations, X is the studied random variable, $\bar{X}$ is the mean of X, and w_{ij} is the weight between X_i and X_j. When $I > 0$ there is positive autocorrelation and when $I < 0$ negative autocorrelation. We aim to calculate the correlation between the number of people that visits a cell and its surrounding cells with the time. Therefore, in our context, N is the number of cells whose the

Table 4: Classes and some of their categories

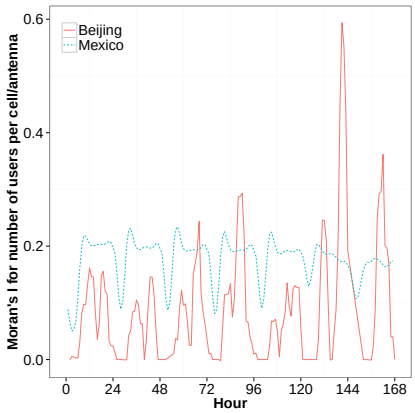
Class	Abbr.	Categories
Arts & Entertainment	A/E	aquarium, casino, etc.
Education	Edu	school, university, etc.
Food	Food	cafe, restaurant, etc.
Religion	Rel	church, mosque, etc.
Outdoor & Sports	O/S	gym, stadium, etc.
Night Life	NL	bar and night club
Shopping	Shop	book store, shopping mall, etc.
Travel	Trvl	bus station, subway station, etc.
Services	Srvc	atm, dentist, doctor, etc.

city was divided into, X is a cell, w_{ij} is the inverse of the distance between cell centers, X_i is the number of people in cell i, and $\bar{X}$ is the mean of people that visited all cells. To aggregate the temporal aspect on the spatial correlation, we calculate I on snapshots of 1 hour, i. e., we sum up the number of people that visited all the cells during one hour and calculate I. Figure 3.9(c) shows the hourly Moran's I for the number of users per cell during one week, from 10th to 16th November, 2008 in Beijing and for the number of subscribers per antenna in Mexico. In order to remove noise in the plot curves, it has been smoothed with a sliding window of size 4 hours applying the average. It is interesting to see the periodical behavior that matches with the diurnal activities on both cities. It means that people tend to crowd some popular areas and its neighborhoods on certain hours of the day. It is particularly true around lunch time on metropolitan areas when people go to common areas of restaurants. Moreover, the right end of the curve represents the weekend on both curves, in which the autocorrelation is higher for Beijing and slightly lower on Mexico. The difference lays in the nature of both datasets. First, Mexico dataset has slightly less users making calls during weekends, i.e., there are less people sharing the antennas, thus spatiotemporal correlation is lower than on weekdays. On the other hand, on a mobility dataset as GeoLife, people increase their mobility during weekends, and, as consequence they gather on common leisure areas more than weekdays.



(a)

(b)



(c)

Figure 9: (Better seen in colors) Coverage time provided by each of the PIs class per (a) period of the day and (b) per weekday and weekends in Beijing. (c) Hourly Moran's I for the number of users on the cells (Beijing) and connected to the antennas (Mexico).

3.4 CONCLUSIONS

In this chapter we have made an extensive analysis of human mobility on several cities in order to unveil common aspects present in the human mobility. Section 3.2.1 presented our system model, which unifies different datasets into a common representation of urban scenario. Then, Section 3.3.1 presented analysis on the visiting patterns to PIs. Results unveiled a clear repetitivity on people's visiting behavior. Additionally, we have proposed a metric to measure how repetitive are the visits of people to PIs. Next, in Section 3.3.2 we have evaluated displacement in people's trajectories. Two are the main conclusions, people have a tendency to use shortest-path when moving around and the mobility is confined, i. e., displacement is generally limited to 10 km. Finally, in Section 3.3.3 we have zoomed out from the per-user analysis to a spatial autocorrelation. It shows that the regular patterns in the human mobility are not restricted to the scale of the dataset, GeoLife and Telco datasets are orders of magnitude distant on the number of users. Moreover, they represent mobility in different ways and in different granularities. Still, their spatial autocorrelation shows the routinary regular behavior. In the following chapter we investigate how routinary aspect is present on the traffic behavior.

4 CONTENT ANALYSIS

"Men's natures are alike; it is their habits that separate them."

— Confucius

4.1 INTRODUCTION

Smartphone devices provide today the best means of gathering users information about content consumption behavior on a large scale. In this context, the literature is rich in work studying and modeling users mobility, but little is publicly known about users content consumption patterns. The *understanding of users' mobile data traffic demands* is of fundamental importance when looking for solutions to manage the recent boost up of mobile data usage [4, 78, 54] and to improve the quality of communication service provided, favoring the proliferation of pervasive communication. Hence, the definition of a *usage pattern* can allow telecommunication operators to better forseen future demanded traffic and consequently, to better (1) deploy data offloading hotspots or (2) timely plan network resources allocation and then, set subscription plans.

Contrarily to most related work in the literature modeling call traffic using CDRs, we characterize and model real mobile data traffic demands generated by smartphone subscribers. Although convenient and of frequent consideration, call records only provide an intuition of users activity in the network: voice calls and SMS. In addition, due to its sparsity in time [44], subscribers behavior in terms of call shows strong variations with time and day of the week [54]: a different behavior is found when data traffic is considered. Finally, call traffic does not describe the background traffic load automatically generated by current smartphone applications (e. g., email checks, synchronization). We thus claim that, since smartphones are now used more for data than for calls [79], the use of call records for investigating traffic demands is not enough for dimensioning network usages.

Our first contribution in this chapter is *to profile urban mobile data traffic*. For this, we perform a *precise characterization of individual subscribers' traffic behavior clustered by their usage patterns*, instead of a network-wide data traffic view [62, 57, 64]. Note that the high dynamic behavior of individual subscribers (in terms of traffic demands and in time) and the use of large scale datasets make this task complex. In addition, for the purpose of quality of service testing of new applications, infrastructures, or network mechanisms, one needs a traffic generator that is capable of generating realistic synthetic traffic

that "looks like" traffic found on an actual network. In this context, our second contribution is *to provide a way for synthetically, still consistently, reproducing usage patterns of mobile subscribers* – the first work in the literature to do so, to the best of our knowledge. The implications of this work are diverse, in particular, in resource allocation planning and testing, or hotspot deployment. When it comes to legal issues, it is also worth mentioning the unconstrained utility of the generated synthetic datasets in practice: synthetic datasets bring no privacy issues to subscribers, and may be used by any entity willing to perform realistic network simulations.

Our study is performed on an anonymized dataset collected at the core of a major 3G network of Mexico's capital (Section 4.2). The dataset spans 4 months from July to October 2013 and consists of all data traffic associated with 6.8 million subscribers. The dataset describes detailed information on the volume and frequency of any data traffic generated by smartphone subscribers. This includes any uploaded and downloaded data traffic, i. e., not only browsing or SMS traffic, but traffic automatically generated by applications are also included. This represents an order of thousands of Terabytes exchanged in the biggest city of Mexico. Moreover, the dataset provides information about age and gender for more than half million subscribers.

We focus on the temporal dynamics of individual subscriber's usage pattern. Thus, we first analyse their traffic usage habits as a function of time, age, and gender (Section 4.2). We observe identical usage patterns on different days. This motivates us to choose one day for studying the subscribers' usage pattern (i. e., "when" and "how much" traffic is generated) in detail. Then, in order to be able to consistently analyse the usage heterogeneity of a larger number of subscribers, we classify them into six distinct profiles according to their usage pattern (Section 4.3). We finally model the usage pattern of these six subscriber profiles according to two different journey periods: peak and non-peak hours. Using a sample and numerous statistical tools, we show the effectiveness of our traffic modeling, which is capable of consistently imitating different subscribers profiles in two journey periods, when compared to the original traffic dataset (Section 4.4). Our main outcome is *a synthetic measurement-based mobile data traffic generator, capable of imitating traffic-related activity patterns of six different categories of subscribers, during two time periods of a routinary normal day in their lives.* We discuss implications of our contributions in Section 4.5. Finally, Section 4.6 concludes this chaper. In this chapter, the words *user* and *subscriber* will be used interchangeably.

4.2 DATASET

The final goal of our work is a measurement-driven traffic modeling. The traffic modeling is performed after several measurement-driven analysis of an anonymized dataset provided by a major cellular operator in Mexico. This dataset captures subscribers' traffic activities generated by 6.8 million smartphone devices located within the large urban area of Mexico city. The data includes information about subscribers' *sessions* that took place from 1st July to 31st October, 2013. It is important to highlight the concept of a session in our work. In the 3G standards, 3GPP or 3GPP2, a session is created when the radio channel is allocated to a subscriber as soon as he has data to be sent. Radio channel might be seen generically as a radio resource, e. g., time slot, code, or frequency. The session is finished by the network after a period of dormancy presented by the subscriber, which is configurable and typically set from 5 to 30 seconds [80]. The studied dataset contains more than 1 billion sessions and each of them has the following information fields: (1) amount of upload and download volumes (in KiloBytes) during the session; (2) session duration in seconds; and (3) timestamp indicating when the session starts.

Furthermore, due to a special characteristic of this dataset, information about age and gender is available for 548,000 subscribers. This allows us to investigate the interesting relation between users' age, gender, and network traffic demands, which can be used by telecommunication operators to better set subscription plans.

Due to the routinary behavior of people [78] and the large scale dataset, it suffices to study a subset of the whole dataset in order to capture the daily behavior of subscribers. Indeed, our analysis shows that there is low variability on subscribers' activity among the same hours on different days. Therefore, we have selected one week to more deeply assess the subscribers' behavior. The studied week spans from 25th August to 31st August 2013 and contains information of about 2.8 million smartphone devices (the highest number of devices among the dataset weeks) and activity that totalizes 104 million sessions. This week has no special days or holidays and it is out of the Mexican preferred vacation period, which spans from early July to mid-August. From the data contained in this week, we have seen an enormous frequency of outliers on the first hour of all days, likely generated by the probe when the data collection was done. Therefore, we have discarded data from midnight to 1am of all days in the following analysis. This does not affect our methodology since it is indifferent to the amount of valid hours that the dataset provides.

Selecting a subset of one week allows us to better assess the subscribers' behavior but it is important to emphasize that we will use the whole dataset later to evaluate our mobile traffic generator. Moreover, contrarily to datasets only describing CDRs, the richness of the

considered dataset allows us to study and to model detailed and realistic data traffic demands over time.

In the following, we study the behavior of mobile subscribers in terms of traffic they generate. The analysis are performed according to four main traffic parameters: number of sessions, inter-arrival time (referred as IAT, the difference between the arrival timestamps of subsequent sessions of the same subscriber), session duration, and volume of traffic.

4.2.1 *Traffic dynamics*

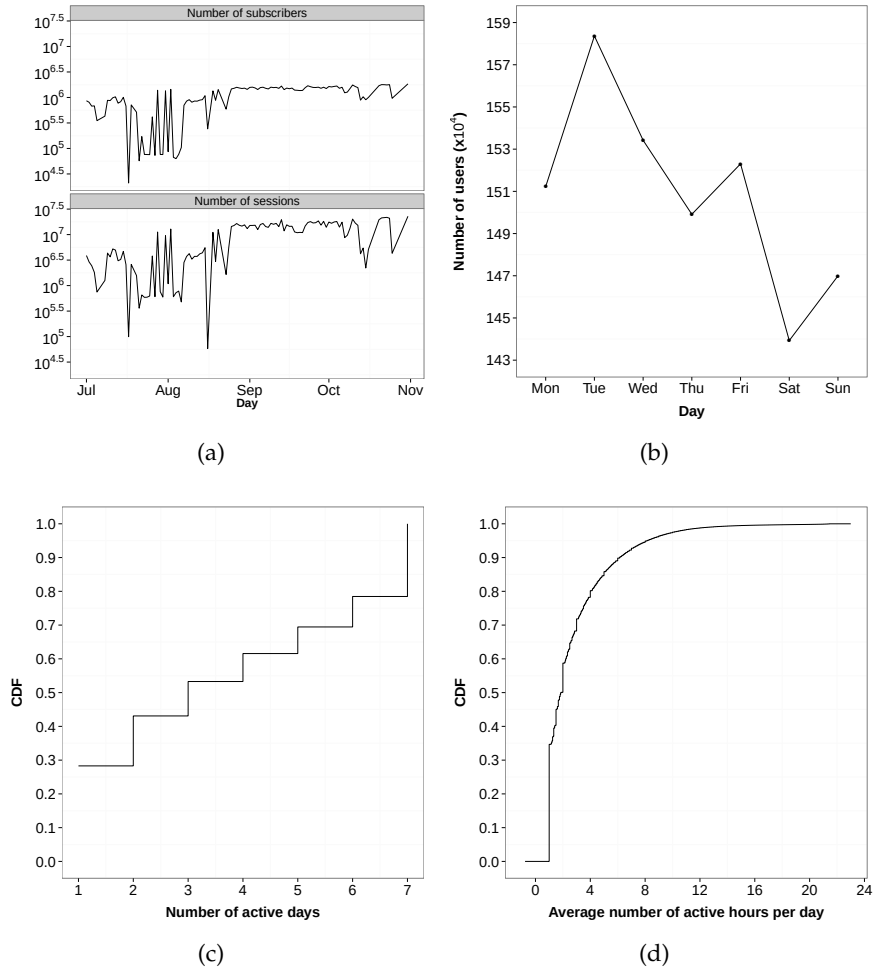


Figure 10: (a) Number of subscribers and sessions on the whole dataset. (b) Number of subscribers per day generating traffic. (c) CDF of number of days in which subscribers generate traffic. (d) CDF of number of hours in which subscribers generate traffic per day during the week.

Figure 4.10(a) shows the total number of subscribers and the total number of sessions from the whole dataset. As expected, *the number*

of subscribers and number of sessions are highly correlated. It is possible to see a similarity on the shape of the curves for both parameters. Indeed, Spearman's correlation between number of users and number of sessions is 98%.

In Figure 4.10(b), we present the number of subscribers that generated traffic on each of the days throughout the selected week (recall that the selected week is 25th August to 31st August 2013). *The day-wise number of active subscribers is essentially decreasing as the week progresses.* The difference between the weekdays and the weekend in terms of active subscribers is considerable; the highest difference is 10% which is obtained by comparing Tuesday with Saturday. As expected, *on average, the number of active subscribers are higher during the weekdays than during the weekend* (also observed in [81]). In the studied week, this average difference is 5%.

Figure 4.10(c) shows the CDF (Cumulative Distribution Function) of the number of active days of the subscribers within the week (a subscriber is said to be active on some day if she generates some traffic on that day).

It is interesting to see that 22% of the subscribers generated traffic on all days, while 29% of the subscribers generated traffic only on one day of the week. Also, 53% of the subscribers generated traffic on three or less days during the week. Similar percentages were measured from a different dataset and reported in [81].

Similarly, in Figure 4.10(d), we show the CDF of the average number of active hours of the subscribers per day. We see that *most of subscribers generate traffic on few hours during the day.* Indeed, on an average 80% of the subscribers generate traffic for up to 4 hours each day. If we consider a longer period, e. g., for up to 6 hours, the number of such subscribers reaches 90%.

Figure 4.11(d) shows the total number of sessions per user per day of the week. *There is a slightly less amount of sessions per user during weekends and a general similarity between the cumulative values for all days.* For instance, considering users with up to 10 sessions per day, the difference between the number of sessions per user on weekdays and weekends is 4%, and 0.1% considering up to 100 sessions per day.

Figure 4.11(a) presents the CDF of session duration per subscriber during the week. *We see a median usage of 63 seconds of session and a significant variation in the duration length of sessions.* Interestingly, most of the sessions present short duration and few subscribers (less than 1%) use more than 6 hours of session during the week. In particular, the duration of 58% of the sessions is at most 100 seconds, while 90% of the sessions lasts for up to 15 minutes (similar behavior was reported in [81]).

Figure 4.11(b) shows the CDFs of the average upload and download volumes of traffic generated per session. Observe that both the upload and download CDFs are similar: e. g., 35% and 38% of the sessions,

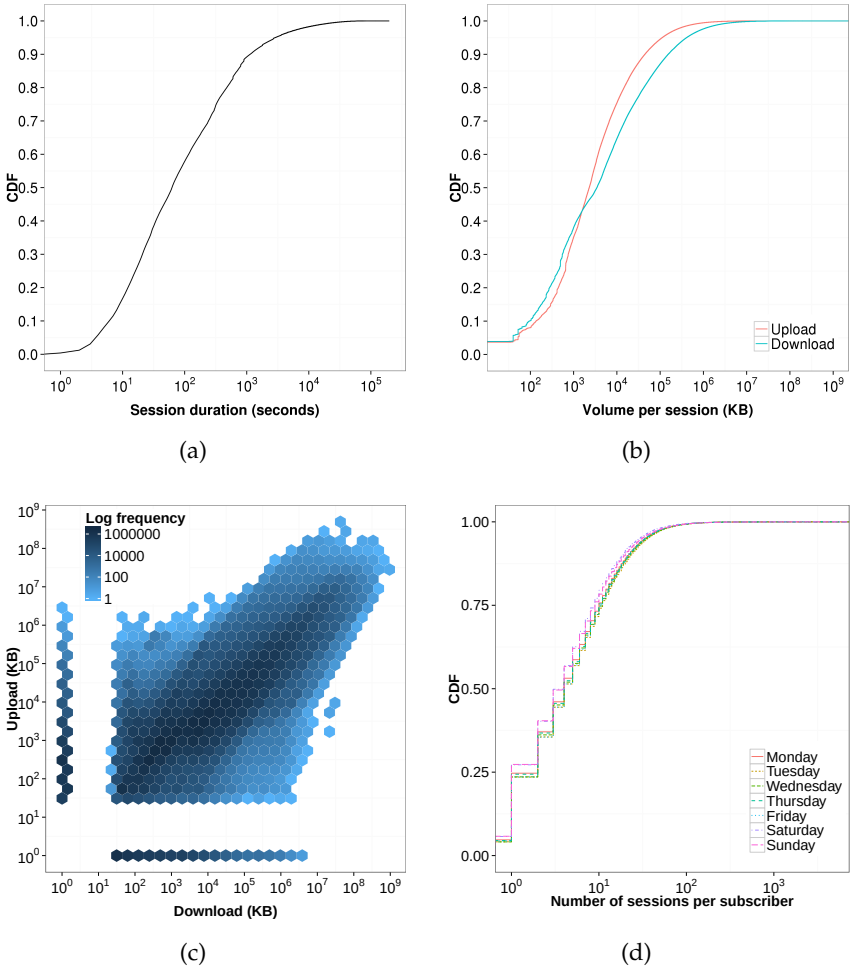


Figure 11: (a) CDF of session duration in seconds per subscriber during the week. (b) CDF and (c) bin plot of the upload and download volume during the week. (d) Number of session per subscriber per day of the week.

respectively, present upload and download volume of up to 1 MB. On the other hand, 6% and 13% of the sessions present more than 100 MB for uploaded and downloaded volume, respectively. *We observe that the median traffic load generated by typical subscribers is not significant while there are a small number of "heavy hitters" that consume a significant amount of network resources.*

Figure 4.11(c) shows the *hexagonal bin plot* [76] of uploaded and downloaded volumes per session during the week. The intensity of a bin represents the frequency of sessions that generated upload and download volumes laying within the bin. *The hexagonal bin plot reveals an uphill pattern from left to right, indicating a positive linear relationship between the per-session uploaded and downloaded volumes.* That is, if the amount of downloaded traffic is higher in a session, we can expect the uploaded volume to be higher as well. Indeed, the Spearman's corre-

lation coefficient between per-session upload and download traffic is 88%. We also observe two groups of bins forming straight lines, one close to each of the axis. Bins close to the x-axis are due to sessions that present a small upload volume, e. g., around 1 KB, and significantly higher amount of download. Those are likely sessions in which subscribers use streaming media sites, e. g., Youtube, that typically use Real Time Protocol (RTP). RTP does not require the subscribers' device to generate confirmation packets, which justifies the small amount of uploaded volume. On the other hand, bins close to the y-axis represent sessions with small amount of download and comparably higher amount of upload. That is probably due to upload of media formats, e. g., photos on Facebook or videos on Youtube.

Owing to the high correlation between the upload and download volumes, in our evaluation and traffic modeling, we take into consideration the total volume per session, i. e., the sum of the upload and download volumes during the session.

4.2.2 *Temporal dynamics*

It is common knowledge that some hours tend to be more active than others when it comes to users routinary daily activities. In this context, peak hours present high frequency of requests and volume of traffic, while non-peak hours present less traffic demands and volume. Indeed, Figures 4.12(a), 4.12(b) and 4.12(c) show three parameters and their hourly dynamics during the week. Two features are important to highlight: *First, there is a repetitive behavior during different days at the same hours. Second, there are peak and non-peak hours when it comes to subscribers' traffic demands.* In the following, we discuss these features and measure how repetitive their behavior is. We further develop the idea of peak and non-peak hours for the users' activity in our traffic model.

Figure 4.12(a) shows the average number of sessions per subscriber on each hour during the studied week. The results show a clear gap on the average number of sessions from 4am to 8am. *On the end of late night and beginning of the day subscribers tend to perform less sessions*. This is consistent with diurnal human activity patterns. The number of sessions generated from 4am to 8am is 10% less when compared with that generated during the rest of the day. Furthermore, the total number of sessions from 9am to 3am is 47% higher than from 4am to 8am. Such behavior repeats over all days of the week.

Figure 4.12(b) shows the upload and download session volumes per user during the week. *Similar to the number of sessions behavior (Figure 4.12(a)), it is possible to see both: the gap between 4am to 8am and the day-wise similarity*.

Figure 4.12(c) shows the inter-arrival time (IAT) of subsequent sessions of the same subscriber. The high IAT shown from 4am to 8am

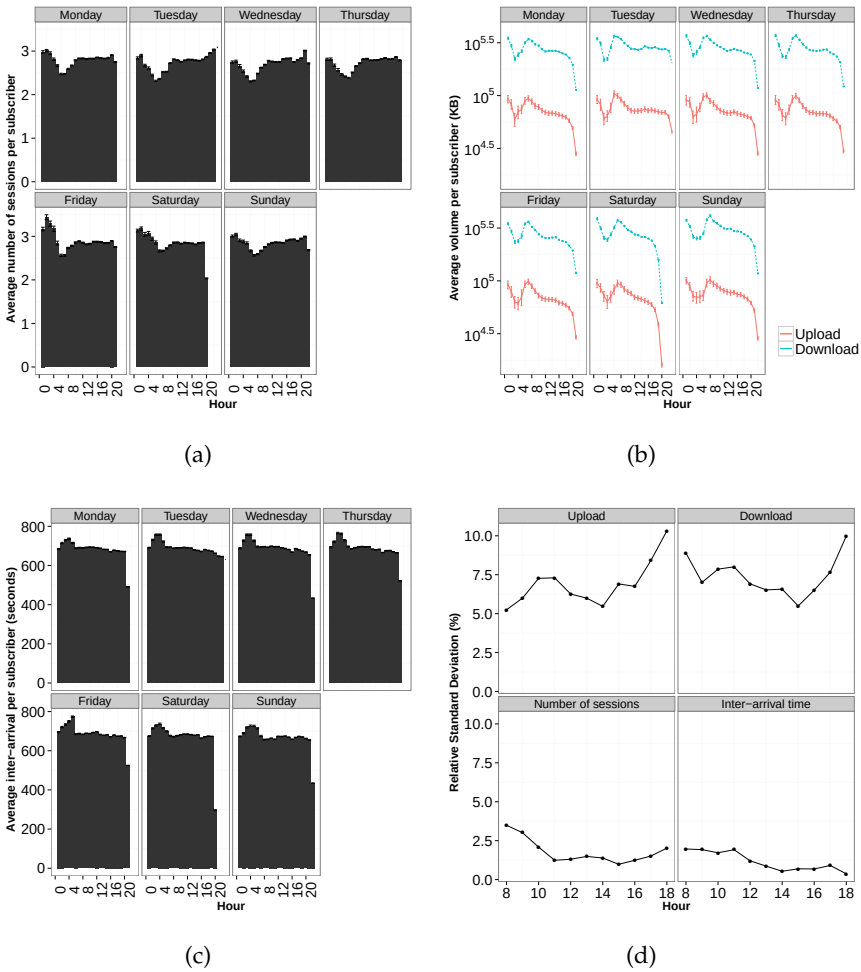


Figure 12: (a) Average number of sessions per user during the week. (b) Volume of traffic for upload and download during the week. (c) Inter-arrival time per subscriber during the week. (d) Relative Standard Deviation per parameter.

is a complementary behavior to the low average number of sessions on the same hours present in Figure 4.12(a). This is expected and due to the fact that *longer inter-arrival times results in less number of sessions on average*.

In summary, these last three results show *a high day-wise similarity on number of sessions, volume of traffic, and inter-arrival time traffic parameters. Indeed, all traffic parameters have similar per-hour values on different days, even comparing weekdays and weekends*. We measure the day-wise variability on subscribers' behavior using the Relative Standard Deviation (RSD). RSD is the absolute value of the coefficient of variation (CV), which is defined as the ratio of the standard deviation σ to the mean μ. Figure 4.12(d) shows the per-parameter average RSD, which considers the hour-wise variation from all 7 days during Mexican working hours (i. e., from 8am to 6pm). It is calculated us-

ing the values of the parameters of the same hours for all the days, e. g., the RSD for the number of sessions at 10 a.m. among all days is 2.08%. It is possible to see that the maximum variability is small for all parameters: 3.4% for number of sessions, 1.9% for inter-arrival time, 10.3% and 9.9% for upload and download volumes, respectively. In order to show that the variability within the day is higher than the variability among the same hours on different days, we have calculated the maximum RSD of each parameter on all hours of each day. For instance, the variability for the uploaded volume on all hours on Friday is 12.7%. The results shows that, on average, 4% for number of sessions, 2% for inter-arrival time, 16% and 15% for upload and download volumes, respectively. Therefore, we can conclude that, *on the studied dataset, the parameters from the same hours on different days present less variability than the parameters within the same day on different hours*.

Contrarily to our findings, previous related studies considering phone records (or CDRs) [54] show that subscribers behavior in terms of call traffic have strong variation with time and day of the week. Instead, our results show the consideration of real data traffic (instead of call traffic) (1) reveal a different facet of subscribers behavior and (2) stress the imprecisions brought by CDRs analysis to the resource allocation planning.

The similarity of the temporal activity patterns among different days of the week is due to people's natural routinary behavior. Therefore, *we select one day (28th August 2013, a Wednesday) of the week to perform our extensive per-hour analysis and distinguish users profiles.*

4.2.3 *Age and gender dynamics*

Among the 2.8 million subscribers in the week mentioned in Section 4.2.1, a subset of 548 thousand of them present personal information regarding age and gender. All analysis in this section refer to this subset of users. Thus, to better understand how age and gender impacts traffic demands, hereafter, we present our analysis on the traffic parameters when considering these new social information.

As any study considering social aspects of participating entities, it is important to understand in which cultural context the measurements are made. Similarly to many Latin American countries, Mexican culture presents gender wage gap that disfavours women [82]. Consequently, having less purchasing power the Mexican women consume less goods. As a probable consequence, from almost half million users of the considered dataset, 56% are men and 44% are women.

Figure 4.13(a) depicts the population pyramid grouped by age and gender. This graphic shows the frequency of age and genders' occurrences with females shown on the left and males on the right. *Regardless of the gender, it is possible to see a higher number of subscribers with*

age range from 25 to 34 years old. Indeed, 33% of the subscribers fall in this range.

To ease the understanding of the per-age behavior, we have defined 4 age ranges: [15, 24], [25, 34], [35, 49] and [50, 85], i. e., users younger than 25, from 25 to 34, from 35 to 49, and over 50 years old. Users younger than 15 and older than 85 years old were removed from the trace. In effect, the small amount of users in those two groups make it difficult to draw any statistical conclusion about them. Figure 4.13(b) shows the percentage of subscribers grouped by gender and age ranges. It is possible to see a higher percentual of male (and consequently less female) users in all age ranges. An interesting aspect of this graphic is the increasing gap between the genders as the age range progresses. To undercover this aspect we have plotted Figure 4.13(c). It shows the percentage of users per age and gender. It is interesting to see that the gap increases with increasing age. The Spearman's correlation between age and age percentage per gender is 87% per male and, consequently, -87% per female, i. e., in our dataset the *male participation percentually increases as the user age increases. Conversely, the female participation decreases with the increase of the age.*

Figure 4.13(d) shows the percentage of active users per age and day. An interesting aspect in this graphic is shown for Saturday and Sunday, that have different age range activities when compared to the rest of the days of the week: The absence of the gap present on weekdays from 4 am to 8 am for users within the [25, 34] range, i. e., an activity growth for subscribers from 25 to 34 years old. This is probably due to the nightly activities that usually attracts younger people on weekends, e. g., bars and night clubs.

Figure 4.14(a) shows the frequency of the number of sessions performed by the subscribers grouped by their ages in all days of the week. In order to improve its visualisation, it does not display the few occurrences in which the number of sessions surpassed 500. Still, it depicts 99.99% of the data related to subscribers' number of sessions. Similar to the day-wise similarity presented in Section 4.2.2, this graphic shows that the age-wise number of sessions is similar on different days of the studied week. Regardless of the age, most of users present low and similar number of sessions per day (see Figure 4.11(d)). Briefly, *per age behavior shows that younger subscribers tend to have peak number of sessions that are higher than older subscribers.*

Figure 4.14(b) better shows the decreasing behavior of the traffic parameters with the increase of the age regardless of the gender. It depicts the mean of four traffic parameters by user grouped per age and gender. As there are few users older than 70 years old, their mean values tends to be noisy. If we consider users up to 70 years old, there is a high negative correlation between age and each of traffic parameters for males and females, respectively, -96% and -95% with volume of traffic, -85% and -71% with number of sessions and -63%

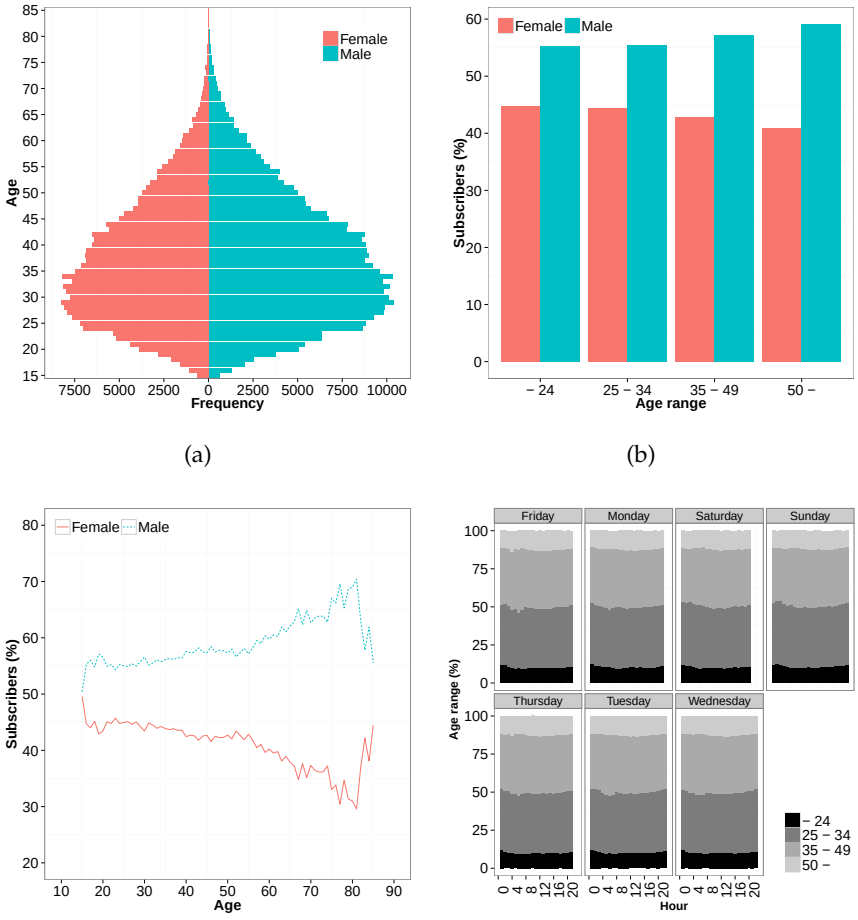


Figure 13: (Better seen in colors) (a) Population pyramid grouped by age and gender. (b) Subscribers by gender per age ranges. (c) Percentage of active users by age. (d) Percentage of active users by age range.

and -78% with session duration. It means that as the age grows, the value of each of those traffic parameters decrease. Except from the inter-arrival time, there is a clear gap between the maximum and the minimum values for each of the parameters from younger to older subscribers, mainly regarding the total volume of traffic. In order to measure this difference, we have calculated the fraction of the traffic parameters from the oldest age range divided by the youngest one. Indeed, *users from the youngest age range generate, on average, 52% more traffic volume, 21% more sessions, 12% longer sessions with the same inter-arrival time*. Generally speaking, in our dataset *users' network activity tend to decrease with the increase of their age.* Our analysis also show the same decreasing activity when subscribers are grouped by their genders, i. e., *it is related to the age of the subscribers and not a behavior of a specific gender*.

Figure 4.14(c) and 4.14(d) show the CDF of number of sessions and CDF of session duration, respectively, grouped by age range and subscribers' gender. As already discussed, the mean network demands is higher for younger users than for older users. Grouping users by age range diminishes this gap when compared to the per-age analysis, but allows us to see the cumulative differences. For both genders, (1) 80% of the subscribers of the oldest age range and 76% of the youngest age range generate up to 10 sessions during the day and (2) 48% of the subscribers of the oldest age range and 43% of the youngest age range generate sessions up to 15 minutes during the day. *In summary, our analysis shows that similar number of sessions and session duration results are seen when users are grouped by age range, irrespective of the subscribers gender.*

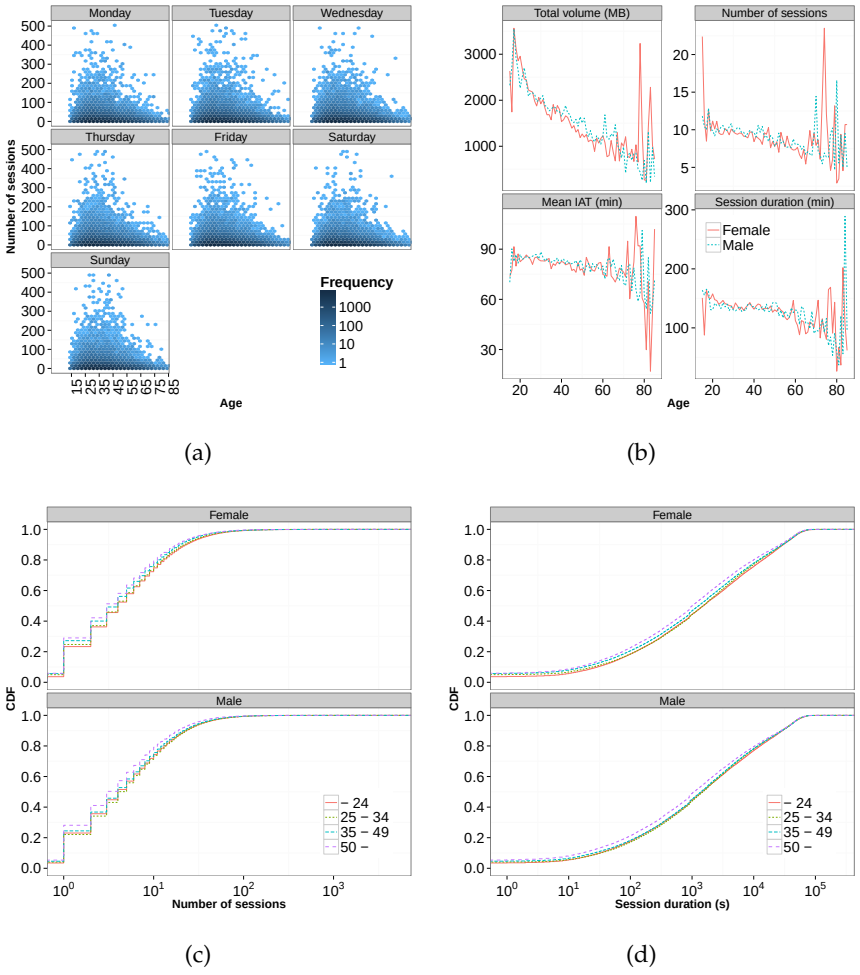


Figure 14: (Better seen in colors) (a) Frequency of sessions per age and day. (b) Mean metrics per age and gender. (c) CDF of the number of sessions per age range and gender. (d) CDF of the session duration per age and gender.

4.3 SUBSCRIBER PROFILING METHODOLOGY

Although having their own repeated routine, human behavior in terms of content demand is highly heterogeneous, as many other human activities. While some subscribers rarely generate mobile data traffic, others demand a few or even a large amount of gigabytes each day. To analyse such different levels of activity, we group subscribers into a limited number of profiles. The profiles are defined according to two traffic parameters, i. e., traffic demands (i. e., volume of traffic) and activity behavior (i. e., number of sessions). Such parameters are extracted from a sample set of the considered dataset describing subscribers' traffic demands. The profile definition is performed in three phases. First, the similarity metric between all pairs of subscribers on a subscribers' sample set is measured according to the two traffic parameters. Second, subscribers are clustered by their similarity into a limited number of clusters, also representing profiles. The third phase allows to classify the remaining additional subscribers of the dataset into the previously defined profiles. This profiling procedure results in typologies of subscribers based on their traffic dynamics. These different phases are detailed in the remainder of this section.

4.3.1 *Similarity computation*

Although we later evaluate our methodologies for a day within the week, our development in this section can hold in general for any time interval D chosen from the week. For a given time interval D, let $\mathbb{S}$ be the set of all subscribers that generate some traffic during D, and $\mathbb{S}' \subseteq \mathbb{S}$ be a randomly selected sample of subscribers from $\mathbb{S}$. Our objective is to partition the subscribers in $\mathbb{S}'$ into a set of *clusters* $\mathbb{P}$, such that subscribers belonging to the same cluster are "similar" in terms of traffic demands. We use Euclidean distance to measure the *similarity* between two subscribers [83]. We then *classify* the remaining users in $\mathbb{S}$ (i. e., $\mathbb{S} - \mathbb{S}'$) into various clusters in $\mathbb{P}$. In this work, we develop a similarity comparison according to *volume of traffic* and *number of sessions*. These traffic parameters allow us to make a comparison between two different subscribers behavior and will be considered at the clustering and classification procedures (discussed in the next section).

Each subscriber i can be effectively represented by the sequence of sessions generated by i. Let t_k^i denote the time instant at which the k-th session of subscriber i begins. Let v_k^i be the volume of traffic (both upload and download) generated by subscriber i during the k-th session. However, this very fine grained representation of a subscriber is costly in terms of memory and processing time required. To overcome this drawback, we divide D into time slots of length T. Thus,

there are $\frac{D}{T}$ number of time slots. The notion of time slots allow us to collect together all sessions occurring within t.

For subscriber $i \in \mathbb{S}'$, let τ_t^i denote the set of all sessions starting within time slot t, i. e., $\tau_t^i = \{k : (t-1)T \leqslant t_k^i \leqslant tT\}$. Now, the volume of traffic generated by subscriber i, in time slot t, is given by

$$V_t^i = \sum_{k \in \tau_t^i} v_k^i. \tag{4}$$

Similarly, the number of sessions generated by subscriber i in time slot t can be written as

$$N_t^i \quad = \quad \sum_k \mathbb{I}(k \in \tau_t^i), \tag{5}$$

where $\mathbb{I}(k \in \tau_t^i) = 1$ if $k \in \tau_t^i$; 0 otherwise. Thus, to obtain N_t^i we simply count the sessions of subscriber i that begin inside time slot t.

Using the above expressions, it is now easy to obtain the total volume and the total number of sessions generated by subscriber i during D: $\vartheta^i = \sum_{t \in D} V_t^i$ and $\eta^i = \sum_{t \in D} N_t^i$. Finally, we define the *traffic volume similarity* between two subscribers i and j as the difference between the total volumes generated by these users, i. e.,

$$w_{ij}^{\vartheta} = \|\vartheta^i - \vartheta^j\|. \tag{6}$$

The *number of sessions similarity* can be similarly defined:

$$w_{ij}^{\eta} = \|\eta^i - \eta^j\|. \tag{7}$$

Using the subscribers in $\mathbb{S}'$ as the vertices, and using either $w_{i,j}^{\vartheta}$ or $w_{i,j}^{\eta}$ as the edge weights, we obtain a complete graph $G(\mathbb{S}', \mathbb{E})$, which is given as input to our clustering algorithm to obtain different clusters in $\mathbb{P}$. The remaining users (i. e., $\mathbb{S} - \mathbb{S}'$) are then classified into the previous defined clusters.

4.3.2 *Subscriber clustering and classification*

Instead of a-priori fixing a value for the number of profiles (i. e., clusters) $|\mathbb{P}|$, our goal is to obtain from the data, how many profiles are needed to best represent the subscribers' traffic activities. For this purpose, we use an hierarchical clustering algorithm that iteratively aggregates vertices from the similarity graph $G(\mathbb{S}', \mathbb{E})$ into larger clusters, according to a dendogram structure [84]. The hierarchical clustering algorithm we choose is the *Average Linkage clustering method*, also known as *Unweighted Pair Group Method with Arithmetic Mean (UPGMA)* [84].

Recall we first group a sample set of $|\mathbb{S}'|$ subscribers into $|\mathbb{P}|$ clusters. Then, we classify the remaining $|\mathbb{S} - \mathbb{S}'|$ subscribers into $\mathbb{P}$. Thus,

UPGMA starts by first considering each vertex of the given graph $G(\mathbb{S}', \mathbb{E})$ as a cluster (i. e., singleton clusters). At each iteration, it computes the distance (using the edge weights between vertices given by Eq. (6) or Eq. (7)) between all pairs of clusters, and then merges the closest two clusters. In our context, it merges together the two clusters that are more similar in terms of traffic demands. If the algorithm is not stopped, it finally simply yields a single cluster containing all the vertices.

Thus, it is important to find where UPGMA should stop its merging process, yielding the best number of clusters, i. e., *the best separation among the groups of usage pattern from subscribers*. To that end, we use several *stopping rules* (or stopping criteria). A stopping rule, during each iteration of the hierarchical clustering algorithm (or each level of the dendogram), gives a measure of how well separated the clusters are, based on which one can decide the best number of clusters to use.

In the literature, there are several stopping rules [85]. Contrarily to related works that have implemented and applied very few of them [81] and in order to avoid to be biased by a specific criteria, we have implemented and used 23 stopping rules, namely Ball-Hall, Beale, Cubic Clustering Criterion, Calinski-Harabasz, C-index, DB, Duda, Dunn, Frey, Friedman, Hartigan, Krzanowski-Lai, Marriot, McClain-Rao, Pseudot2, Ratkowsky-Lance, Rubin, Scott-Symons, SDbw, SD, Silhouette, TraceW, TraceCovW [85, 86, 87, 83, 88, 89].

For the sake of illustration, we will briefly describe the C-Index [85] stopping rule here. C-Index is defined as $C = (S - S_{min})/(S_{max} - S_{min})$, where: (1) S is the sum of all distances between pairs of users in the same cluster over all clusters, (2) S_{min} and S_{max} are the sum of the smallest and the largest distances respectively, for all pairs of users, over all clusters. In our context, it compares the distances among the considered traffic parameters. According to C-Index, the lower the value of the index, the better the clustering. In this way, the number of profiles producing the lowest C-Index value is the one that grants the best separation among clusters.

Figure 4.15(a) shows the C-Index index values as a function of the number of clusters, when number of sessions similarity is considered at the distance computation between pairs of users. C-Index considers choosing the best number of clusters based on its minimum index value. Thus, the best number of clusters is 2 according to Figure 4.15(a).

Similarly, each other 21 implemented stopping rules listed above define their best number of clusters to be used. In Figure 4.15(b), we present the frequency of the best number of clusters, while profiling subscribers using traffic volume similarity. It condensates in a histogram the result of the 23 stopping rules. It shows that 8 stopping

rules recommend 3 as the best profiles, when clustering subscribers by their traffic volumes.

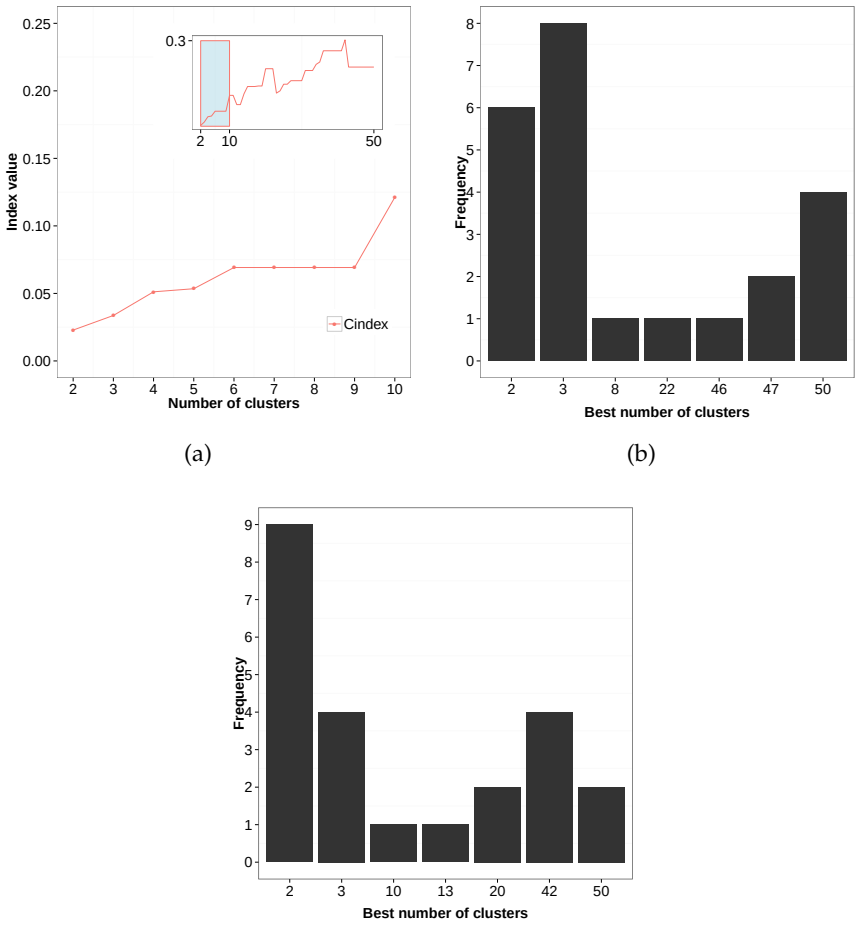


Figure 15: (a) C-Index values and respective number of clusters when re-clustering subscribers at the 3rd defined "traffic-volume"-based cluster, according to the number of sessions similarity. (b) Histogram of best number of "traffic-volume"-based clusters indicated by the assessed stopping rules. (c) Histogram of best number of "number of sessions"-based clusters indicated, when re-clustering subscribers at the 2nd defined "traffic-volume"-based cluster.

Profiling occurs then in four stages: (1) building a similarity graph with $|\mathbb{S}'|$ subscribers, (2) hierarchically clustering it using a similarity metric, (3) determining the best number of clusters $|\mathbb{P}|$, i. e., profiles relying on the stopping rules, and (4) classifying $|\mathbb{S} - \mathbb{S}'|$ remaining unclassified subscribers in the previous defined clusters.

In the fourth stage, we use the *k-means algorithm* as the classification technique. It is worth mentioning, we calculate the clusters centroids (means) obtained from the hierarchical clusters and use them on the first iteration of the k-means algorithm. This is an important informa-

tion because the centroids obtained from the hierarchical clustering algorithm are likely to be better positioned than the k-means originally bootstrapped initial centroids, which are based on randomly selected positions.

These four stages are performed in two rounds. In the first round, the graph $G(\mathbb{S}', \mathbb{E})$ weighted according to the *traffic volume similarity* (Eq. (6)) is used at the hierarchical clustering. The best number of "traffic volume"-based clusters is then determined: according to the results shown in Figure 4.15(b), $|\mathbb{P}| = 3$ weighted subgraphs $\{G_1(\mathbb{S}'_1, \mathbb{E}), G_2(\mathbb{S}'_2, \mathbb{E})\}, G_3(\mathbb{S}'_3, \mathbb{E})\}$ are created. At the end of the first round, the final classification of $|\mathbb{S} - \mathbb{S}'|$ subscribers takes place. The next execution round initiates with a new hierarchical clustering being performed inside each initially defined "traffic volume"-based cluster. This time G_1, G_2 and G_3 are weighted according to the *number of sessions similarity* (Eq. (7)). Finally, for each of these three initial clusters, two "number of sessions"-based clusters are defined after the second round of stopping rules execution (e. g., Figure 4.15(c)), totalizing six subscribers profiles. Due to space constraints, we will not show all stopping rules results. The second round ends with the classification of the remaining $|\mathbb{S} - \mathbb{S}'|$ subscribers into the six defined profiles. Next section better details our subscriber profiling.

4.3.3 *Subscriber profiles*

To obtain the profiles for our dataset, we set D as 27th of August, which contains information of about 1.5 million smartphone devices, and randomly sampled 10000 subscribers (thus, $|\mathbb{S}'| = 10000$ to be used in the clustering procedure). D is a normal day with no special event or holiday and we divide it into time slots of duration T. Time slots help to understand the general behavior of a certain period of time in D. Higher the number of time slots, shorter is their duration and vice-versa. Very short time slots, e. g., 1 minute, may lead to an analysis with fewer sessions per time slot, hindering the identification of subscribers' behavior per slot. Very large time slots, e. g., 12 hours, may lead to a general view of the sessions, so that it is difficult to obtain a good quality assessment of the traffic dynamics. Thus, for our evaluation, we choose a "moderate" value of 1 hour as the time slot duration. Nevertheless, the optimal size of the time slot is still an open problem [90].

Our profiling methodology resulted in *six profiles*, and we have named them as follows: Light Occasional (LO), Light Frequent (LF), Medium Occasional (MO), Medium Frequent (MF), Heavy Occasional (HO) and Heavy Frequent (HF). *Light* profiles contain subscribers that generate up to 17 MB of data during the day, *Medium* profiles have subscribers that generate between 17 MB and 560 MB of traffic during the day, and *Heavy* profiles contain users that generate more than

560 MB of traffic during the day. Likewise, *Occasional* profiles contain subscribers that generate less connection sessions, whereas *Frequent* profiles contain users generating more connections per day. Tables 5, 6, and 7 show the characteristics of each of the profiles.

Table 5: Characteristics of the Light profile

	Light	
Volume	29 KB to 17305 KB (≈ 17 MB)	
Nº of subscribers	418843	
	Occasional	Frequent
Nº of sessions	1 to 10	11 to 224
Nº of subscribers	405848	12995

Table 6: Characteristics of the Medium profile

	Medium	
Volume	17306 KB to 560044 KB (≈ 560 MB)	
Nº of subscribers	610917	
	Occasional	Frequent
Nº of sessions	1 to 51	52 to 1926
Nº of subscribers	598340	12577

Table 7: Characteristics of the Heavy profile

	Heavy	
Volume	560046 KB to 655769309 KB (≈ 650 GB)	
Nº of subscribers	487141	
	Occasional	Frequent
Nº of sessions	1 to 316	317 to 8737
Nº of subscribers	484959	2182

In Figure 16, we show the dynamics of the traffic parameters per subscribers' class per hour. Figure 4.16(a), 4.16(b), and 4.16(c) corresponds to the number of sessions, volume of traffic, and the mean inter-arrival time, respectively; the error bars correspond to a 95% confidence interval. For each time slot, the volume of traffic and number of sessions are calculated using Eq. (4) and Eq. (5), respectively.

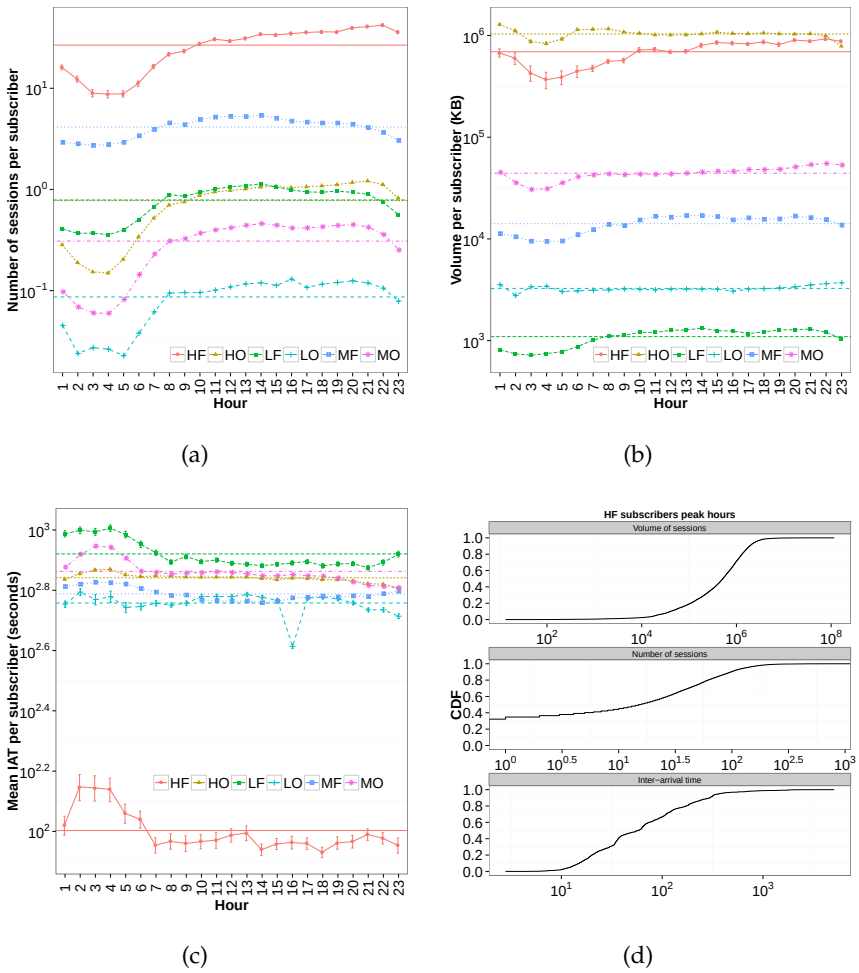


Figure 16: (Better seen in colors) (a) Mean inter-arrival per class. (b) Number of sessions per class. (c) Volume of traffic per class. (d) Empirical CDFs of HF users in peak hours.

For each subscriber i, the average inter-arrival time in time slot t is obtained using the following expression:

$$IAT_t^i = \frac{\sum_{k \in \tau_t^i} (t_{k+1}^i - t_k^i)}{N_t^i}, \tag{8}$$

where τ_t^i is the set of all sessions of subscriber i that lie with the time slot t. Similar to ϑ^i and η^i, we define the average inter-arrival time for the entire D as $\zeta^i = \sum_{t \in D} IAT_i^t$.

From Figure 16, we can see that our methodology well separates the profiles, i. e., the *occasional* and *frequent* subscribers have their values clearly separated. Note that an aggregated traffic analysis would not allow us to identify and consequently, to imitate the behavior of very light users. In fact, the traffic generated by very heavy users (representing a very small percentage of users in the dataset) would bias the analysis and the synthetic traffic generation.

For each curve in Figure 4.16(a), 4.16(b), and 4.16(c), we have also shown a *horizontal line that represents the respective mean value* (where the mean is taken over all time slots). Given the mean values, we classify, for each profile of subscribers and for each parameter (number of sessions, traffic volume, and IAT), the hours above the mean as *peak hours*, and hours below the mean as *non-peak hours*.

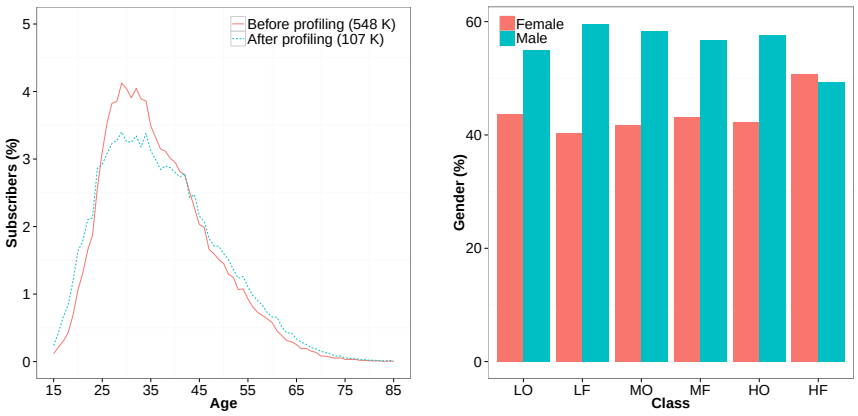
4.3.4 *Profile's age and gender*

In this section, each of the resulting profiles is assessed by the age and gender of their members. The profiled day D has 1.5 million users, from which 107 thousand have information regarding age and gender. The results shown in this section refer to this subset that counts with 57.6% of male and 42.4% of female users. This subset is consistent with the distribution of users with available age and gender prior to the profiling process, which counted 548 thousand subscribers over a week (Section 4.2.3). To evaluate this consistency, we calculate the percentage of users per age on the 548 thousand non profiled users and on the 107 thousand profiled users. Figure 4.17(a) shows this percentage for each of them. There is a visual similarity between the shape of the two curves as they are strongly correlated, with 99% Spearman's correlation.

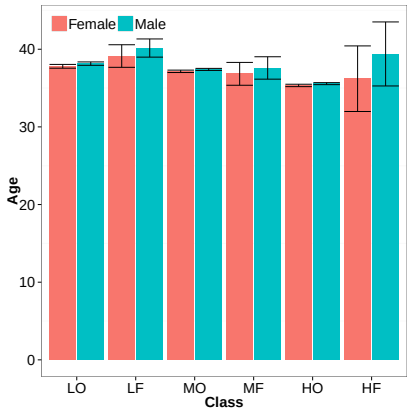
Figure 4.17(b) shows the percentage of male and female subscribers per class, after the profiling of 107 thousand subscribers. *Most of the classes present higher percentual of male than female, except HF in which female have 1% more users than male.* On average, Light and Medium profiles have 15% more males than females, while Heavy profiles have 6% more male than female.

Figure 4.17(c) shows the average subscribers' ages per gender and classes. Due to the large overlapping presented by the confidence intervals (95%), we can assert that *the per-class ages are not significantly different*. That is interesting because *it indicates that the profiles group together users from a wide spectrum of different ages.*

Figure 4.18(a) and 4.18(c) show the CDFs of number of sessions per subscriber per class. The former groups subscribers per age range and the latter per gender. An interesting difference between Occasional and Frequent users is steepness of the CDF curves. *Number of session from Occasional profiles is more uniformly distributed than from Frequent users, which has a very steep slope. It means that most of the Frequent users generate the lowest amount of sessions within the range of their profiles* (recall that the ranges are specified in Tables 5, 6, and 7). *For all classes, male users generate, on average and median, more sessions than females.* On Occasional classes the difference is 1% at most, while on Frequent classes the difference ranges from 2% to 19%. The cumulative values show the same results, for instance the third quartile is



(a) (b)



(c)

Figure 17: (Better seen in colors) (a) Percentage of subscribers per age before and after profiling. (b) Percentage of subscribers per gender and class. (c) Average subscribers' age per gender and class.

at most 1% higher for male than female on all Occasional and LF profiles. Moreover, it is 10% higher on MF and HF profiles.

Figure 4.18(b) depicts the CDFs of session duration per subscribers' class and age range. On average, profiles do not present statistically different session duration values for each of the age ranges. For instance, the per-class confidence intervals (95%) for each of the age ranges overlap each other by the mean. It means, *the session duration behavior within each of the profiles for a certain age range is not statistically different from the behavior of another age within the same profile.*

Figure 4.18(d) presents the kernel density estimation (KDE) curves for the volume per user per gender and class. *There is a similar behavior for male and female subscribers for all the profiles, except HF.* HF male subscribers density curve is narrower than the the female one and present a peak around 10 GB. On the other hand, HF female subscribers curve is wider. It means that, among the heavy and fre-

quent subscribers, male present less diverse session volumes when compared to female.

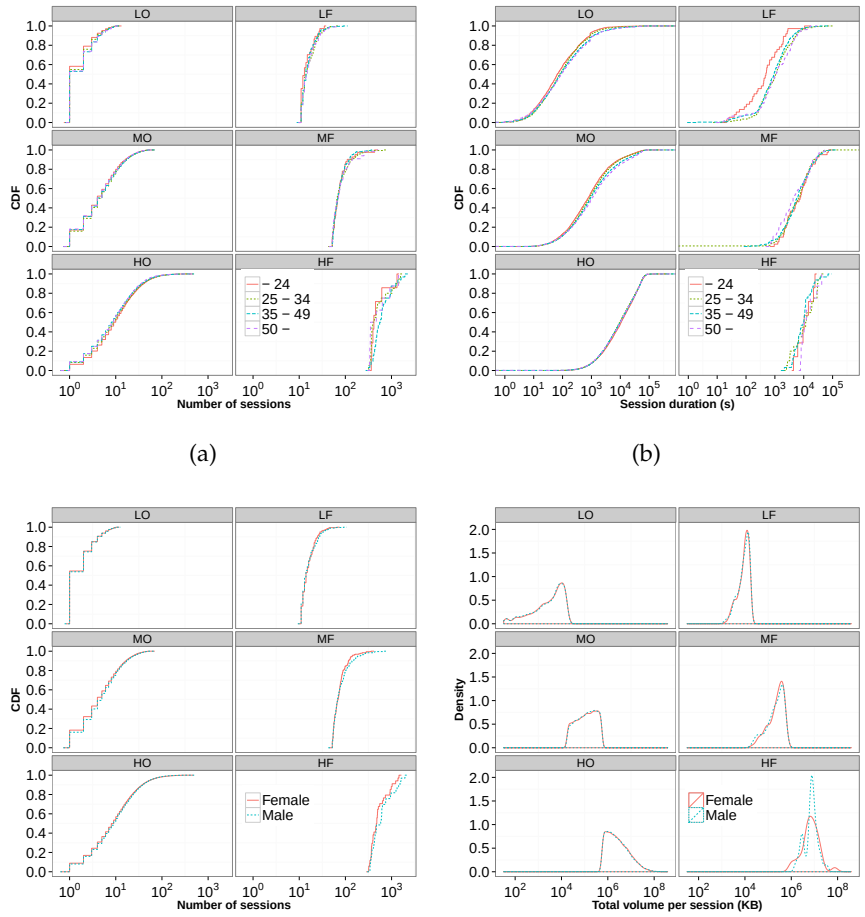


Figure 18: (Better seen in colors) (a) CDFs of number of sessions and (b) session duration per subscribers' class and age range. (c) CDFs of number of sessions and (b) session volume per subscribers' class and gender.

4.4 MEASUREMENT-DRIVEN TRAFFIC MODELING

Realistic network simulations requires a traffic generator capable of imitating actual daily subscribers traffic demands, i. e., has to be consistent with the observations made about the real subscribers in the previous section. Recall that subscribers belonging to different profiles (LO, LF, MO, MF, HO, and HF) have their own specificities in terms of *when* the sessions are generated during the day, and the *volume* generated during each session. Furthermore, each profile of subscribers have different behavior during *peak and non-peak hours*. Thus, to obtain a fine grained model it is important to take into account all the above considerations, while simulating a synthetic trace. In the following, we describe how we merge all the above considerations to obtain a measurement-driven mobile data traffic modeling.

4.4.1 *Fitting empirical distributions*

Using the original subscribers' data, we first study for each profile in peak and non-peak hour, the empirical distribution functions (i. e., CDF) of the traffic parameters (e. g., Figure 4.16(d)): the number of sessions generated, the traffic volume associated with each of these sessions, and the inter-arrival times between the sessions. For instance, the empirical distribution function of *"total volume for HF users in peak hours"* is obtained from the set of all V_t^i (Eq. (4)) such that $i \in S$ is an HF subscriber and t is a peak hour. The empirical distribution functions of the number of sessions and the inter-arrival time for any combination of profile and hour-type (peak or non-peak), can be similarly generated using N_t^i (Eq. (5)) and IAT_t^i (Eq. (8)), respectively. Refer to Section A.2 for the CDFs of the traffic parameters in peak and non-peak hours for all the profiles.

Once the CDFs are obtained, using statistical tests, we estimate the set of distributions that best fit them. From this set, we then select the closest distribution function to the respective CDF. *This function will be used at the traffic usage pattern generation for the corresponding profile and type of hour*. More specifically, when considering the volume of traffic and the inter-arrival time parameters (i. e., consisting of continuous values) of a certain profile and hour, the Kolmogorov-Smirnov statistic test [91] is used. The test estimates the parameters for a set of continuous distributions (namely, Log-normal, Gamma, Weibull, Logis, and Exponential) that best fit the corresponding empirical distribution function. Similarly, when considering the number of sessions parameter (i. e., consisting of discrete values) of a certain profile and hour, the Chi-squared statistic test [92] is used to estimate the best fitting parameters for a set of discrete distributions (Negative binomial, Geometric, and Poisson). In both cases, after getting the sets resulted

from the fitting tests, we select the distribution functions that best fit each corresponding CDF.

Tables 8, 9, and 10 list the best fitted distribution functions along with their parameters for all possible combinations of profile and hour-type pair, for number of sessions, traffic volume and inter-arrival time parameters, respectively. For Negative-binomial distribution, n is the size parameter and p is the probability parameter. For Gamma distribution, α indicates the shape parameter and β is the rate parameter. For Weibull distribution, k is the shape parameter and λ refers to the scale parameter. For Log-normal distribution, σ represents the shape parameter and μ is the scale parameter. For Gamma, Weibull and Log-normal, x_0 is the location parameter.

4.4.2 *Synthetic subscriber generation*

Generating a synthetic subscriber will first require us to generate a profile type (LO, LF, MO, MF, HO, or HF) for the subscriber. Profile types are assigned randomly, based on the distribution of profiles population observed in the real data. For instance, from Table 5, we see that 26.7% of the subscribers belong to LO profile, and thus with probability $q_{LO} = 0.267$ we assign LO profile to a synthetic user. Similarly, the probabilities of other profiles are: $q_{LF} = 0.0085$, $q_{MO} = 0.394$, $q_{MF} = 0.0082$, $q_{HO} = 0.319$, and $q_{HF} = 0.001$. We will refer to $q = (q_{LO}, q_{LF}, q_{MO}, q_{MF}, q_{HO}, q_{HF})$ as the *profile pmf*, or probability mass function.

We now briefly describe our procedure for generating a synthetic subscriber (for a detailed algorithm, refer to Section A.3). *We first randomly generate a profile type for a subscriber* i *using the profile pmf* q. *After obtaining the profile type, for a given hour* t, *we randomly sample values for each traffic parameter according to the corresponding fitted distribution functions*.

In more detail, the algorithm requires one parameter which is the number of synthetic users to be generated (line 1). The result of the generation is a list of sessions per user (line 28). Each synthetic user session contains two fields: (1) volume of traffic and (2) arrival timestamp. For each subscriber i and time slot t, we sample a number of sessions N_t^i, an average session volume V_t^i, and a mean inter-arrival time IAT_t^i from the appropriate distributions (i. e., the fitted distribution corresponding to the profile and hour-type pair) listed in Tables 8, 9, and 10, respectively (lines 5 to 7). The volume per session v_k^i (for $k \in \tau_t^i$, see Section 4.4) is then equal to the sampled value V_t^i divided by the sampled number of sessions N_t^i. The initial timestamp of each session in hour t is then computed according to the sampled inter-arrival time IAT_t^i and number of session N_t^i for that hour (lines 13 to 20). By varying t over the 24 hours in a day, we obtain a synthetic subscriber traffic for one day.

Table 8: Number of sessions: distributions and parameters

Number of sessions			
Hour	*Profile*	*Distribution*	*Parameters*
Peak	HO	Neg-binomial	$n = 0.1139$, $p = 0.09$
	HF		$n = 0.4703$, $p = 0.01$
	MO		$n = 0.1772$, $p = 0.3$
	MF		$n = 0.7588$, $p = 0.13$
	LO		$n = 0.1885$, $p = 0.62$
	LF		$n = 0.4802$, $p = 0.32$
Non-Peak	HO	Neg-binomial	$n = 0.0448$, $p = 0.1$
	HF		$n = 0.1437$, $p = 0.01$
	MO		$n = 0.0536$, $p = 0.3$
	MF		$n = 0.3146$, $p = 0.08$
	LO		$n = 0.0810$, $p = 0.66$
	LF		$n = 0.2405$, $p = 0.33$

4.4.3 *Synthetic traffic model evaluation*

In order to evaluate our traffic modeling, we generate a synthetic dataset and compare it with the original dataset. Towards this goal, we first generate a set $\mathbb{R}$ of synthetic subscribers, where $|\mathbb{R}| = |S|$, for one day of traffic. The synthetic dataset contains for each session of a subscriber i and at hour t: (1) the volume in KiloBytes generated and (2) the initial timestamp of the session.

Let $\mathbb{D}$ denote a set of different time periods including D and the synthetic day denoted as D'. $\mathbb{D}$ also contains each day from 1st July to 31st October, i. e., the whole dataset. Let p^e_ϑ denote the PDF (Probability Distribution Function) of the total volume generated by a subscriber active in day e in the original trace, formally defined as $p^e_\vartheta(x) = \sum_{i \in e} \mathbb{I}(\vartheta^i = x)/|\{i \in e\}|$. For a visual comparison, Figure 4.19(a) depicts the CDFs corresponding to the PDFs p^D_ϑ and $p^{D'}_\vartheta$ of traffic generated in the original day D and synthetic day D'. We can observe an *almost complete overlap of the two CDFs due to high similarity between the real trace and the synthetic trace.*

We then assess, how consistent the synthetic traffic is by comparing the distributions of the various parameters between the original and the synthetic datasets. For this, we use the Bhattacharyya (BH) measure [93]. It quantifies the similarity between two discrete or continuous probability distributions. Let $p(i)$ and $p'(i)$ be two pmfs, i. e., $\sum_{i=1}^{N} p(i) = \sum_{i=1}^{N} p'(i) = 1$. The BH measure is formally defined as $\rho(p, p') = \sum_{i=1}^{N} \sqrt{p(i)p'(i)}$. However, the BH measure is not a

Table 9: Session volume: distributions and parameters

Session volume			
Hour	*Profile*	*Distribution*	*Parameters*
Peak	HO	Weibull	$k = 0.49$, $\lambda = 476551.7$, $x_0 = 30$
	HF		$k = 0.81$, $\lambda = 774639.6$, $x_0 = 40$
	MO		$k = 0.59$, $\lambda = 31936.8$, $x_0 = 29$
	MF		$k = 0.80$, $\lambda = 13959.4$, $x_0 = 37$
	LO		$k = 0.85$, $\lambda = 3228.7$, $x_0 = 29$
	LF		$k = 0.92$, $\lambda = 1181.7$, $x_0 = 33$
Non-Peak	HO	Weibull	$k = 0.50$, $\lambda = 452332.8$, $x_0 = 30$
	HF		$k = 0.63$, $\lambda = 384935.6$, $x_0 = 40$
	MO		$k = 0.58$, $\lambda = 26617.7$, $x_0 = 30$
	MF		$k = 0.79$, $\lambda = 10657.9$, $x_0 = 33$
	LO		$k = 0.79$, $\lambda = 2800.1$, $x_0 = 29$
	LF		$k = 1.03$, $\lambda = 873.5$, $x_0 = 34$

distance metric since it does not satisfy all the metric axioms. Therefore, [94] proposes an alternative distance metric based on the BH measure which is formally defined as $d(p, p') = \sqrt{1 - \rho(p, p')}$. Note that, $d(p, p')$ exists for all discrete distributions and it is equal to zero if and only if $p = p'$. We use d in order to measure the similarity between the original dataset and the synthetic dataset. Literature provides other metrics to calculate the distance between two probability distributions, e.g., Kullback-leibler (KL) divergence [95], which is well-known in the Information Theory field. In our context, KL diverge is not suitable, since it does not allow the divergence to be calculated on vectors of different sizes. Since we compare how similar are the traffic demands among different days, it is unlikely that all the days contain the same number of subscribers. On the contrary, BH works regardless the sizes of the vectors that form the CDFs and that is the main motivation on usage of BH instead of KL in our scenario.

We first compute $d(p_\vartheta^D, p_\vartheta^{D'})$, the distance between the total volume distribution of the original day and the synthetic day. Then, we compute $d(p_\vartheta^D, p_\vartheta^e)$, $e \in \mathbb{D}$ but $e \neq D$, the distance between the original day and remaining days in the original trace. We obtain similar distances for p_η^e and p_ζ^e for $e \in \mathbb{D}$, which are respectively, the PDFs of the total number of sessions and average inter-arrival time by a subscriber active in day e. Finally, for each distribution, we have also computed the mean and the confidence interval (95%) of the distances between the original day and the remaining days. In Figure 4.19(b), we show the $d(p_\vartheta^D, p_\vartheta^e)$ distances (cf. $d(p_\eta^D, p_\eta^e)$ and $d(p_\zeta^D, p_\zeta^e)$). Also

Table 10: Session mean inter-arrival times: distributions and parameters

Session mean inter-arrival time			
Hour	*Profile*	*Distribution*	*Parameters*
Peak	HO	Gamma	$\alpha = 1.2517$, $\beta = 0.0017$, $x_0 = 0.5$
	HF	Log-normal	$\sigma = 4.0917$, $\mu = 1.1285$, $x_0 = 4.68$
	MO	Gamma	$\alpha = 1.2990$, $\beta = 0.0016$, $x_0 = 0.5$
	MF	Gamma	$\alpha = 2.2081$, $\beta = 0.0034$, $x_0 = 1$
	LO	Weibull	$k = 0.8508$, $\lambda = 548.24$, $x_0 = 1$
	LF	Gamma	$\alpha = 1.7929$, $\beta = 0.0019$, $x_0 = 2$
Non-Peak	HO	Gamma	$\alpha = 1.2044$, $\beta = 0.0017$, $x_0 = 0.5$
	HF	Log-normal	$\sigma = 3.9374$, $\mu = 0.9822$, $x_0 = 3$
	MO	Gamma	$\alpha = 1.1921$, $\beta = 0.0017$, $x_0 = 0.5$
	MF	Gamma	$\alpha = 2.0301$, $\beta = 0.0034$, $x_0 = 1$
	LO	Gamma	$\alpha = 0.7078$, $\beta = 0.0013$, $x_0 = 1$
	LF	Weibull	$k = 1.1988$, $\lambda = 827.96$, $x_0 = 1$

shown in Figure 4.19(b) (horizontal dashed line) is the $d(p_\vartheta^D, p_\vartheta^{D'})$ distance (cf. $d(p_\eta^D, p_\eta^{D'})$ and $d(p_\zeta^D, p_\zeta^{D'})$). The traffic model evaluation consists then in verifying whether the $d(p_\vartheta^D, p_\vartheta^{D'})$ is within the confidence interval of the $d(p_\vartheta^D, p_\vartheta^e)$. As can be seen in Figure 4.19(b), *for each distribution, the distance of the synthetic day (from the original) is within this confidence interval.*

Finally, we applied the profiling methodology described in Section 4.3 on the synthetic subscribers. By doing so, we classify them and compare the per-class traffic behavior with the one created from the original dataset. Figure 4.19(c) depicts the per-class behavior for the volume of traffic per session for the classified synthetic subscribers. It is possible to see that *this result is coherent with the one for the original dataset* presented in Figure 4.16(b). For instance, the behavior for peak and non-peak hours is well defined and similar to the one from the original trace.

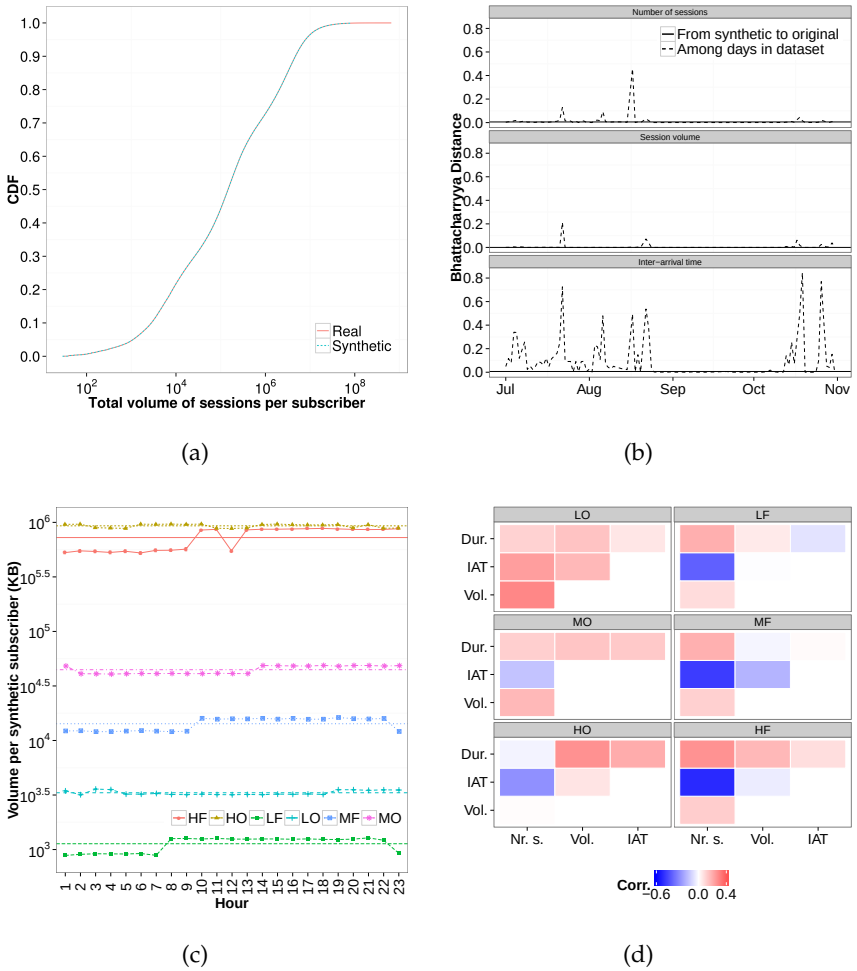


Figure 19: (Better seen in colors) (a) CDF of the total volume generated by real and synthetic subscribers (b) Per-parameter BH distances between original and synthetic trace (dashed line) in D, and between the original trace in D and other days e from the original trace (full line) (c) Volume of traffic per class for synthetic subscribers. (d) Heatmap of the correlation between session duration, inter-arrival time and volume of traffic.

4.5 DISCUSSION

In this section, we discuss some issues we judge interesting in the presented work. An important aspect on network planning and management is to know what is the load it will be subjected to. Subscribers with different profiles impose, on certain cases, totally different demands to the network. For example, our dataset shows that the heaviest user generates 22 million times more traffic than the lightest one. Moreover, the 276 thousand lightest subscribers generate similar amount of traffic as generated by a unique heaviest subscriber in the entire day.

Traffic demand is generally described by the set of different traffic parameters that characterize the demands of the users to the network. In this work, we have explored a set of parameters such as inter-arrival time, session duration, number of sessions, and volume of traffic. Alone, each of those parameters were deeply assessed on our previous sections, but it is also interesting to see what is the relation among them.

Figure 4.19(d) shows a heatmap of the Pearson's correlation between those traffic parameters for all subscribers in all profiles. The intensity of the color on each cell of the matrix indicates how strong is the negative or positive correlation. It is possible to see that the correlation between number of sessions and inter-arrival time goes from a low positive value on LO to a high negative value on HF. Indeed, the correlation between them is 22%, -14% -26% -37% -45% -55% for LO, MO, HO, LF, MF, and HF, respectively. *It means that classes in which subscribers generate more sessions have higher negative correlation with the inter-arrival time. In general, the more sessions a user generates, the shorter they need to be to fit in a certain period of time.* A caveat here is that a user that generates few sessions could generate them in bursts, or sparsely separated in time. The former would result in small IAT and the latter in a larger IAT. For example, a large IAT of one hour is likely to be done for a user with few sessions per day, than a user with, for example, 300 sessions. In the same way, a small IAT could be generated for a user with both high or low number of sessions.

Another important aspect is the relation between volume of traffic and session duration. LO, MO, and HO classes present 13%, 14%, and 26%, respectively, i. e., a growing positive correlation with the session duration. LO, MO, and HO have, on average, 663, 6554 and 18624 seconds of session duration and 5090, 165214 and 6117322 KB of average session volume, respectively. *The growth of those metrics from one Occasional class to the next is due to the necessary increase on the session duration in order to accommodate the volume of traffic, considering that there is no significant raise on the number of sessions from LO, to MO, or HO.*

Finally, it is important to mention the correlation between number of sessions and volume of traffic. The correlation is overall low and positive between these two metrics for all profiles, but its behavior differs completely from Occasional to Frequent users. LO, MO, and HO have 29%, 17% and 0.4% correlation between number of sessions and volume of traffic, respectively, i. e., a decrease from LO, to MO and to HO. It happens because LO users have few sessions and low traffic volume, while MO and HO classes have significantly higher volume of traffic, but still few sessions. Therefore, the correlation is lower for MO and HO than for LO. Differently, LF, MF, and HF have 9%, 11% and 12% correlation between number of sessions and volume of traffic, respectively, i. e., a growth from LF, to MF and to HF. That is due to HF presenting both high volume and high number of sessions, while LF and MF present lower volume of traffic, but still high number of sessions.

Understanding network demands from users traffic parameters and their correlations is one of the contributions of our work. Moreover, this work provides distributions to model workload characteristics of mobile subscribers' traffic demands and a framework on how to create a traffic generator out of it. Therefore, it has implication in areas related to the design of new applications and network mechanisms as well as network planning such as hotspot deployment [78].

In this latter area, for instance, the objective is to provide the best placement for hotspots respecting certain constraints. For instance, one may desire to deploy a fixed amount of hotspots to maximize the amount of data offloaded from the network. The literature frequently presents the evaluation of hotspot deployment based on mobility datasets describing subscribers' trajectories. Although literature provides some mobility datasets, to the best of our knowledge none of them provides information of both mobility and traffic demands. Our traffic generator could be attached to the mobility datasets and it would allow to better exploit them. Besides, a synthetic traffic generator allows the generation of traffic demands of any size of population: While the original traffic dataset allow the sampling of users up to the size of the dataset, a synthetic traffic generator allows to expand this limit.

Another important aspect of the synthetic traffic generator is that it preserves the privacy of the original subscribers from whom the measurements came from. The non-existence of personal data attached to synthetic users allows us to limitlessly share our observations with the community without the necessity of sharing sensitive information inherent of datasets. One may argue that it is possible to anonymize the users identity, but literature shows that many attempts on that direction fail on protecting users privacy [96]. As shown in our analysis, our synthetic users generate traffic consistent with the original dataset and, thus do not carry privacy issues.

4.6 CONCLUSIONS

In this chapter we have presented a network traffic, temporal and social characterizations of a 4-month dataset that contains more than 1.05 billion session connections from about 6.8 million smartphone users. In Section 4.2.1, our traffic analysis shows short sessions are predominant and a high correlation between the amount of upload and download. Next, in Section 4.2.2 we show that, for the evaluated dataset, subscribers present more similarity in terms of network traffic among the same hours on different days than within the same day in different hours. Later, Section 4.2.3 presented an interesting age and gender analysis of the subscribers related to their network demands. We have shown that male subscribers are more activity than the female counterparts as the age passes by. Moreover, regardless the gender, the younger people were more active than older ones. In Section 4.3 we propose a framework that automatically classifies those users by their traffic demands into a limited number of profiles. Our approach takes advantage of repetitive user behavior due to their daily routines. Furthermore, we have provided distributions that describe their traffic demands into peak and non-peak hours. Finally, from these distributions, in Section 4.4 we create a traffic generator and evaluate the synthetic trace it generates. Our results showed that the synthetic trace presents a consistent behavior when compared to original dataset.

5 STUDY CASE

"The highest reward for man's toil is not what he gets for it, but what he becomes by it."

— John Ruskin

5.1 INTRODUCTION

With the steady growth of smart-phones sales [97], the demand for services that generate mobile data traffic has grown tremendously. In the last four years, the used amount mobile data has grown more than twice in relation to the previous year. The average generation of mobile data traffic grew 81% in 2012, and the global data traffic will increase 8-fold until 2017 [98]. This recent boost up of mobile data consumption are struggling the 3G cellular networks, which are not always prepared to receive such demand [99]. From 2007 to 2012, AT&T faced 20,000 percent growth in mobile data traffic on its network [10]. In 2009 AT&T saw its network overloaded by mobile data traffic, mainly generated by iPhone users and, as a result, an unresponsive system with delayed text messages, dropped calls, and slow download speed [100].

Wi-Fi offloading seems to be promising solution to the recent boost up of mobile data consumption that is making excessive demands on 3G cellular networks in metropolitan areas. [13]. The idea consists in shifting the traffic off of cellular networks to Wi-Fi networks. Carefully deploying Wi-Fi hotspots can both be cheaper than upgrade the current cellular network structure and can concede significant improvement in the network capacity [13, 14]. Nevertheless, one question remains: *how Wi-Fi hotspots should be deployed?* The following factors make the answer to this question a challenge task.

The expansion of metropolitan areas increased the possibility of moving around [101]. This fact together with the increase of smartphone use results in highly dynamic links, which may significantly affect the performance of the network [102]. Moreover, people may use different transportation modes, which significantly impacts their trajectories: e. g., a person riding a bike or walking can decide the path to follow contrarily to someone inside a bus. Finally, it is also important to take into account the space-time interaction between people and urban locations, a key point for an efficient network planning. Such considerations can reveal fundamental insights in terms of network usability. Popular sites for instance, are the source of the most of the traffic on the network [103].

In order to consider these issues, *this chapter tackles the Wi-Fi hotspot deployment problem in a metropolitan area by leveraging mobile users' context and content, i.e., their trajectories, scenario interaction and traffic demands.* To the best of our knowledge, this is the first work leveraging users' three dimensional information (time, space and interest) in data offloading services. Our objective is to define what are the best places to receive Wi-Fi hotspots in order to maximize the offloaded traffic on an urban scenario. This is a convenient solution for both cellular operators and users: The former can see the traffic being shifted to inexpensive networks while the latter can take advantage of higher data rates and less monetary costs than using cellular networks. We claim that unplanned deployment of hotspots may lead to both under-utilized and over-utilized network areas.

Related to our work, [104, 105] proposes the current approaches in the literature to deploy hotspots that considers user mobility characteristics only. Section 5.2 presents a deep comparison between those two works and the one herein proposed.

To accomplish our objective, we study the mobility context of people in a metropolitan area of a major city and identify a set of locations to well deploy Wi-Fi hotspots (cf. Section 3.2). Our strategy (cf. Section 5.3) is methodologically structured as follows. First, we create a time dependent graph to represent the interaction between people's mobility and locations suitable to receive a hotspot. Then, we measure how much data offloading a location can contribute to. For this, we use a metric herein proposed to rank which locations are suitable to support more data offloading: Better positioned hotspots are likely able to offload more data.

Through experiments on a real-life trace, we evaluate the performance of our routine-based network deployment in terms of offloaded traffic, by varying the number of deployed hotspots. We also compare our solution with the current work on the literature. The results reveal that when using a realistic synthetic traffic model, our strategy provides up to 12% more offloaded traffic than the current solution on the literature (cf. Section 5.4). Finally, Section 5.6 concludes this chapter.

5.2 RELATED WORK

Literature contains works that study access point deployment as a way to alleviate high traffic demands on the mobile networks. [106] proposes offloading models to increase the capacity of large-scale heterogeneous networks, which already contain deployed hotspots. The work herein proposed differs mainly because we focus on the deployment of hotspots, and a possible fine-tuning on their capacity is out of our scope. [107] studies the impact of the indoor Wi-Fi and Femto cell deployment on data offloading for a mobile network. Differently, our work focus is on city-wide outdoor deployment.

In [108] authors evaluated the cost of single and multi-access networks with heterogeneous traffic. The approach consisted of macro cells being deposited to ensure full coverage and, subsequently, micro, pico, or WLAN access points are placed to meet certain specific demands of capacity. Firstly, conclusions indicated that the combination of macro, micro, and pico cells did not achieve lower costs per user when compared with single access approach. After further investigation, authors quantified the infrastructure cost for a multi-access network composed of macro cellular HSDPA base stations and IEEE 802.11g WLAN access points on an urban scenario and concluded that the costs can be lowered considerably by adding a hotspot layer [109, 110].

Moreover, some efforts were made to create deployment strategies focusing on scenarios with one specific mobility mode, e. g., walking, or driving. On [104] authors propose a Wi-Fi deployment strategy that takes in consideration user mobility characteristics on a college campus. The objective is to provide continuous coverage for a mobile user, i. e., maximize the probability that a mobile user can connect with an AP wherever it moves on a given area. The study is conduced using the Dartmouth College campus dataset [111], and based on the physical structure of the campus, a mobile graph is built considering that vertices are buildings or road intersections, and edges as paths between them. Thus, two NP-Hard sub-problems are then studied, the maximum continuous coverage, and the minimum deployment cost. The main issue with this approach is that it considers and evaluates only a campus-wide scenario with only one popular area, while on a urban scenarios many spots are likely to have similar importance. Indeed, it is important to understand the mobility on a restricted environment in which people are mostly walking, but on the other hand, our approach differs on taking an urban scenario with a diverse set of transportation modes.

On [112, 113] authors propose a centrality-based approach do deploy Road Side Units (RSUs) in order to improve performance aspects on both vehicle-to-vehicle and vehicle-to-infrastructure communication. On [114] authors show results on how Femtocells can improve

the quality of a network on a vehicular environment. On [115] authors aim to facilitate the identification of locations that are capable of providing data offload for vehicular networks. To do so, they propose an application that uses different sources of information, e. g., GPS, Wi-Fi beacon footprints, and tower triangulation. Diversely, our work differs when it considers several transportation modes and the inherent different demands they have on the network infra structure.

In [13] a quantitative study is made to asses the performance of mobile data offloading through Wi-Fi networks. The study was conduced based on data collected from 100 users in Seoul. Due battery consumption restrictions, the offloading was not actually executed on the user's phone but simulated based on the collected connectivity patterns. The authors defined offloading efficiency to be the total bytes transferred using Wi-Fi by the total bytes generated. The simulations shows that the on-the-spot offloading, i. e., spontaneous use of Wi-Fi as soon as it is available, can achieve 65% efficiency, increasable in 13% to 21% using delayed offloading, i. e., avoiding 3G traffic while waiting a certain amount of time for a Wi-Fi connection. The aforementioned work does not propose a deployment strategy but concludes that carefully deploying more Wi-Fi hotspots can both be cheaper than upgrade the network structure and can concede substantial improvement in the network capacity.

On [14] authors proposed an architecture to offload data called MobiTribe. The strategy chooses a set of privileged users called *tribe* and shares popularly requested content to be stored on their devices. Whenever a user requests for the content available in the tribe, the choice of the content provider is based on the network connectivity and the load on the available devices in the tribe. Users looking for content request it first to the members of the tribe and if the content is not available, they request it from the Content Manager Server, e. g., Youtube. In order to share the data inside the tribe, users need to wait for its pairs to have access to an inexpensive network interface such as Wi-Fi. Results show that with 1 hour delay, the efficiency of the architecture can reach up to 89% for data upload. This delayed approach focus on scenarios in which the real-time data transmission factor is not crucial.

Similarly, [116] use a set of special nodes called *VIPs* to gather and offload data aiming to avoid the usage of the cellular network on peak periods. As any other node, VIP nodes represent devices being carried by individuals. The strategy applies centrality metrics in order to select the least number of VIP nodes which can cover the rest of the devices in the network. A percentage of the most central nodes composes the VIP set, which it is going to be responsible for gathering the data from the rest of users of the network, and upload it in during non-peak periods of the day. The results show that the centrality-based approach creates, on average, VIP sets with less than 7% of total

nodes while and it is able to offload 90% of the data. Differently, our approach does not put users in charge of the data offloading but aims to better deploy the network to provide more offload opportunities. Nevertheless, VIP strategy and our approach are not mutually exclusive, thus, they could be used in conjunction. On the other hand, VIP scenarios consider walk and vehicular mobility on campus and urban scenarios, but we focus on wider range of transportation modes. Other works such as [117] use similar strategy, i.e., chooses a set of targeted-users to receive and propagate data using local connections, but differs due to the context of a Mobile Social Network. Authors propose data offloading solution as a multi-layer network composed by different cells, typically, macro, micro, pico and femto. This work differs from the herein proposed because it focus how people interact when specifically walking by each other.

To the best of our knowledge, [104, 105] are the most prominent approaches to deploy Wi-Fi that takes in consideration user mobility characteristics. They consist of non-delayed access-point based data offloading strategies as classified in [118]. Data is not stored prior to the offloading, i. e., it goes straight from the users' phone to the operator through an access point.

In [104] authors create a mobility graph in which edges correspond to the road section traveled by the user and nodes are either buildings or road intersections. Weight on the edges represent the number of Access Points (APs) required to completely cover the route. Based on a dataset that indicates which user is connected to which AP, authors define the probability of a path being taken by a person. The probabilities guide the algorithms on choosing which APs are going to be deployed. Differently, in [105] no graph is built. Instead, the city is divided as a grid in which one hotspot will be positioned on each cell center. Based on the mobility of taxis, one data request is made every 5 seconds. Hotspots are ranked by how many data requests they offload, i. e., better hotspots are those positioned in cells with more data requests. Both [104, 105] present some significant differences from the work herein proposed which are summarized in Table 11.

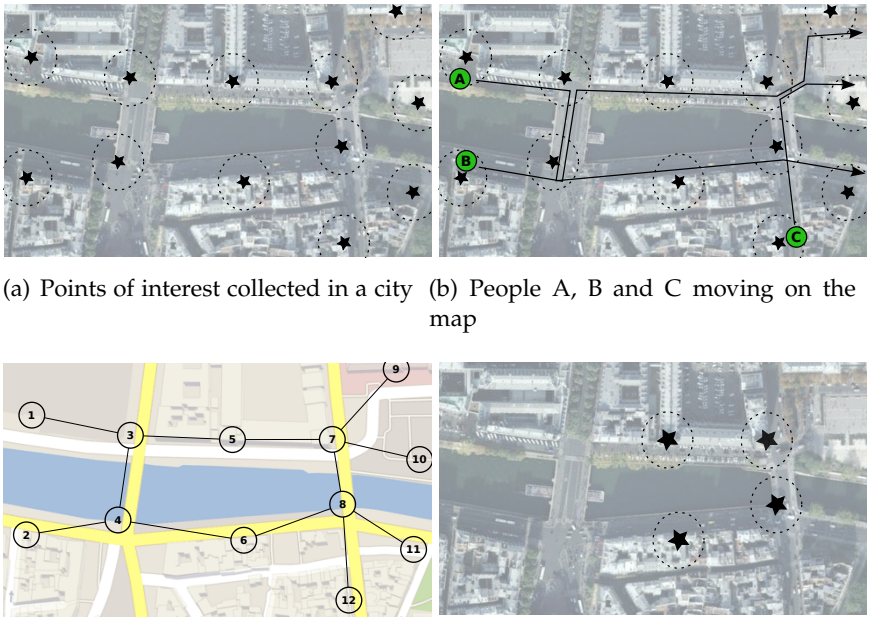
First, our objective is not to provide continuous coverage as considered in [104]: In an urban scenario, this is prohibitively expensive since it would require the deployment of hotspots over the whole area, where most of them may be under-utilized. Similar to [105], our goal is to maximize the offload ratio. Second, we consider an urban scenario, which presents significantly higher complexity than the campus scenario considered in [104]: i. e., higher densities, many popular areas, diverse types of mobility (imposed by a variety of transportation modes), bigger area, etc. [105] employs a metropolitan scenario but considers taxi as the only transportation mode. Third, our approach is not restricted to the consideration of only one popular spot as done in [104]: In an urban scenario, diverse popular areas

Table 11: Related work comparison

Work	[104]	[105]	Our proposal
Objective	Maximize continuous coverage	Maximize offloaded traffic	Maximize offloaded traffic
Scenario	Campus	City	City
Transportation modes	Walking	Taxi	10 different
Points of Interest	Real	Grid	Real
Traffic	Not considered	One request every 5 seconds	Based in the literature

may exist and their features may vary according to space-time issues. Fourth, [104] does not apply a traffic model while [105] employs a non realistic traffic model in order to measure the offloaded traffic. Contrarily, our work uses a more realistic synthetic traffic model that is strongly based on parameters and real measurements from the literature.

Some other related works provide solutions for data offloading but not related to hotspot deployment e. g., delegating the data offloading for people's devices [117, 119].



(a) Points of interest collected in a city

(b) People A, B and C moving on the map

(c) Graph induced by the mobility of A, B and C

(d) The best four hotspots chosen

Figure 20: A general view of our proposed methodology.

5.3 PROPOSAL

In order to understand the relation between people and urban scenario during time, we construct a time dependent graph that describes interaction between people's trajectories (edges) and PIs (vertices). This section presents how mobility is mapped into a time dependent graph and introduces our hotspot placement strategy.

5.3.1 *Graph creation*

Let $p \in P = \{1, 2, \ldots, q\}$ be a geographic coordinate on the trajectory of a person. Let $s \in S = \{1, 2, \ldots, n\}$ be a PI. Both P and S are uniquely identified by their geographical position: $\forall s_{(x,y)} \in S, \forall p_{(x,y)} \in P$, where (x, y) is a pair of geographical coordinates. Since we are using real data traces containing GPS-based positioning, x and y are respectively, latitude and longitude. Figure 5.20(a) illustrates a set of PIs presented as stars on map. Initially, we assign all PIs with a hypothetical range illustrated as dashed circles[9].

Our graph $G(V, E)$ represents the interaction between people and the urban scenario. The graph should provide a simple and easy-to-see representation of people's mobility. While walking, people create trajectories and their interactions with the urban scenario may point out important places to start planning the deployment of a network

9 Unit disk graph model.

infra-structure. Figure 5.20(b) illustrates the three trajectories created by three hypothetical people (A, B, and C) and their respective interactions with PIs. A vertex $v \in V = \{1, 2, \ldots, w\}$ where $w \leqslant q$ is created on the same coordinates of a PI if and only if this latter covers at least one person passing by. A coverage of a person a by PI s with a range r is expressed as $\forall p \in P_a, \exists p \,|\, (p_x - s_x)^2 + (p_y - s_y)^2| \leqslant r^2$, i. e., a is inside s's circular range. Note that a PI that does not respect the condition of minimum coverage is not considered as vertex in the graph.

An edge $e \in E$ is created between two vertices if their corresponding PIs sequentially cover a person during its trajectory. Figure 5.20(c) shows the graph when three people are covered by PIs.

5.3.2 *Metrics*

We model the features present on people's routinary mobility according to three concepts presented on Section 3.3: *Desire lines*, *confinement* and *repetitiveness* of people's mobility. While repetitiveness is computed using homonymous metric introduced in Section 3.3.1, Desire lines and Confinement are described by Stress and Closeness Centrality calculated on the vertices of the graph. Following, related details are provided.

STRESS CENTRALITY: Considering such feature, vertices are ranked with the *stress centrality metric* [120]. Consider a graph $G(V, E)$, the Stress Centrality (st) for a certain vertex $v \in V$ is defined as follow:

$$st(v_i) = \sum_{j=1}^{N} \sum_{k=1}^{j-1} \sigma_{jk}(v_i) \tag{9}$$

where N is the number of vertices on the graph, $\sigma_{jk}(v_i)$ is the number of shortest-paths that contains v_i between the pair of vertices v_j and v_k. Vertices with high stress are those that lay on most of people's shortest routes and may become well positioned hotspots. Vertices with high stress values are likely to be places for hotspot deployment. Indeed, centrality metrics are widely used in the literature, but not yet explored in the problem we are considering.

CLOSENESS CENTRALITY: On our proposal, we consider *Closeness Centrality metric* [121] as a way to measure how a place is close to each others. Closeness is calculated based on the geodesic distance between all pairs of vertices on the network. Consider a graph $G(V, E)$, the Closeness Centrality (cl) for a certain vertex $v_i \in V$ is defined as follow:

$$cl(v_i) = \sum_{k=1}^{N} d(v_i, v_k) \tag{10}$$

where N is the number of vertices on the graph, $i \neq k$ and $d(v_i, v_k)$ is the shortest-path distance between v_i and v_k. It assigns higher values for vertices closer to the rest of the network, that is, on a city those are probably hospitals, and markets, i. e., places planned to be close to most of the people's trajectories.

5.3.3 *Synthetic traffic model*

To the best of our knowledge, there is no freely available dataset describing both people's mobility and their traffic demands on an urban scenario. Therefore, we have created a realistic synthetic traffic model that takes into consideration traffic parameters and measurements from the literature [122, 123, 98, 124]. In addition, it takes in consideration different traffic demands required by people using different transportation modes. The traffic model herein discussed is responsible for simulating the content generation and offload as if it was done by people participating on GeoLife, while performing their trajectories. Table 12 shows the parameters used on the synthetic traffic model.

5.3.3.1 *Communication*

While moving around, a person passes by many PIs and sometimes may stop by. We consider that the interaction between a person and a point of interest lasts as long as the former is inside the interaction range of the latter. The interaction range is defined as the Wi-Fi range on a urban scenario. Taking into account the interferences caused by buildings, vehicles, or any other obstacle, the effective Wi-Fi range in outdoor environment can vary from 5 to 75 meters [115]. We consider a interaction range of 50 meters in our experiments: if a person is, at most, 50 meters from a point of interest they are able to wirelessly communicate and consequently, the latter covers the former. It is important to enhance that initially a point of interest is not compulsory considered as a hotspot but as potential place to receive a Wi-Fi hotspot structure.

5.3.3.2 *Content generation*

Considering projections made in [98], by 2017 two-thirds of the mobile data traffic will be used to transfer video. Therefore, our synthetic traffic model will generate data assuming users generating video-sized files. File size is Gamma distributed based on workload characterization study made on 250 thousand popular videos crawled from Youtube on a five month period [123]. Besides, its shape was taken from study made with more than 2.6 million videos crawled from Youtube during a three month period [122]. Furthermore, it has been

shown that the Internet access times of mobile users are exponentially distributed [13].

The amount of data per day and the transfer rate were estimated from monthly volumes projected in [98].

The synthetic traffic model simulates content creation by generating *loads*. A load contains an initial timestamp indicating when it was generated, the file size, and the final timestamp that is the sum of its initial timestamp and transfer time. We consider a constant transfer rate during the connection between an user's device and a hotspot. Therefore, transfer time is the file size of the load divided by the transfer rate. For each leg of a person's trajectory, our model generates a load and divides it throughout the points of this leg as defined in Section 3.2.1 (Figure 5.21(a)). The division respects the duration of the leg and the available transfer rate, i. e., no content can be generated after the leg was traveled and the load is divided to points limited to the maximum transfer rate.

5.3.3.3 *Content offload*

When a hotspot covers a trajectory point, a content offloading is likely to happen (Figure 5.21(b)). Connection duration is taken into consideration on the data offloading, to do so, there is a 1.83 second gap between the start of a coverage and the availability of the data offloading. This was made aiming to simulate the IP acquisition time [124]. Therefore, after the initial connection gap, the traffic generated will be offloaded from trajectories points to the hotspot (Figure 5.21(b)).

Moreover, the synthetic traffic model was designed to couple with specific demands inherent in a scenario that contains people moving using different kinds of transportation modes. The objective is to create a scenario that is closer to the reality and, to do so, synthetic traffic model should consider only reasonable situations in which someone would be generating traffic depending on how this person is moving around. When a person is covered by a PI, and the transportation mode being used is "taxi", "bus", "walk", "train", "subway", "car" or "boat" the traffic generated by our model is normally made available to be offloaded. Correspondingly, when the person gets covered and uses other transportation modes, e. g., "run", "motorcycle" and "bike", no content will be available for offloading.

We call *tr* the sum of all the content (*traffic*) offloaded by a PI from all trajectory points on its coverage range. Formally we define it as $tr(v_i) = \sum_{p \in P'}^{N} l_p$, where P' is the set of points within the coverage range of v_i and l_p is the load on the trajectory point p.

5.3.4 *Objective formalization*

Consider a graph $G_t(V_t, E_t)$ constructed as described in Section 5.3.1 for a period of time $t \in T = \{1, 2, \ldots, u\}$. Let ϕ be the number of pe-

Table 12: Synthetic traffic model parameters

Parameter	Value
File size	Gamma (shape=2 [122], scale=8.5 MB [123])
Inter-arrival time (IAT)	Exponential [13]
Amount per day	94.54 MB [98]
Transfer rate	3.9 Mbps [98]
Connection time	1.83 second [124]

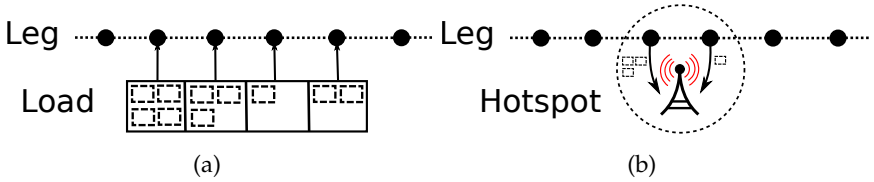


Figure 21: (a) Content generation. (b) Content offload.

riod of times per day $d \in D$, which represent the set of days in which we analyze mobility data. Therefore, $|T| = |D| * \phi$. G_t represents all the interactions between people and urban scenario for the hours contained in t. Therefore, $V_t \subseteq V$ is the set of vertices that covered at least one person during t and $E_t \subseteq E$ is the subset of edges that represents the mobility of people during t. Our objective is to maximize data offloading while limiting the number of deployed hotspots. We advocate that in order to deploy a Wi-Fi network for data offloading on an urban scenario, it is important to take into consideration both traffic and mobility characteristics. For instance, consider two hotspots h_i and h_j that receive a certain amount θ of traffic to be offloaded during a period of time t. If we consider only offloaded traffic, both would be equally selected with no priority. *Our intuition is that if we also consider people's mobility characteristics it is possible to provide a better network deployment for data offloading envisioning future traffic. Popular regions in the cities tend to be more popular and more traffic likely will be generated.*

Consider the mobility characteristics metrics Stress (st), Closeness (cl) and Repetitiveness (re) as described in Section 5.3.2. Consider the traffic (tr) as the content offloaded as described in Section 5.3.3.3. Consider $U_{i,t,m}$ as the value of the metric $m \in M = \{st, cl, re, tr\}$ on the period t for the vertex i. The traffic on a period t for the vertex i is calculated as the sum of all b bytes of content offloaded from all people covered by i during t, i.e. $U_{i,t,tr} = \sum_b b_{i,t}$. Due to the different magnitudes of the metrics, we have normalized $U_{i,t,m}$ between 0 and 1 by the maximum value of m on the period t.

Let $\overrightarrow{R_{i,t}}$ be a vector of metrics calculated for the vertex $i \in V_t$. Due to the mobility, a vertex i may appear on the graph on a certain period $t = k$, but do not appear on a period $t = k + 1$. Therefore, $[R_{i,t,m} = U_{i,t,m} \cdot z_{i,t}]$, where $z_{i,t} = 1$ if i appears on t and $z_{i,t} = 0$ if i does not appear in t. The result of metrics attributed to the same vertex on different periods is aggregated through a sum, i. e., $\sum_t R_{i,t,m} = Total_{m,i} \forall\, m \in M$. Then, our problem is to find the set of vertices that maximizes the metrics with a limited number of hotspots λ, formally:

$$\begin{aligned} &\text{Maximize} \sum_{i,t} R_{i,t,m} \\ &\text{Subject to} \sum z_{i,t} < \lambda \end{aligned} \tag{11}$$

Besides, our model can be tuned in order to prioritize either mobility characteristics or traffic: $\alpha * (\sum_{i,t} R_{i,t,st} + \sum_{i,t} R_{i,t,cl} + \sum_{i,t} R_{i,t,re}) + \beta * \sum_{i,t} R_{i,t,tr}$ by choosing weights α and β depending on operators' demands. Considering $\alpha \to 0$ and $\beta \to 1$ results in a deployment that prioritizes traffic demands. On the other hand, when $\alpha \to 1$ and $\beta \to 0$ hotspots will likely be placed on more important areas of the city regardless their traffic offload. For this first analysis, we are going to consider both traffic and mobility characteristics with the same importance.

5.3.4.1 *Spots selection*

The metrics are calculated on a offline centralized manner, i. e., the solution herein proposed may be applied on data previously collected by the interested entity, e. g., a Telco operator. This premise is reasonable since network deployment planning generally uses historical data as input. The result is a set of PIs ranked by how good they are in providing traffic offloading. Besides, the PIs are greedily selected based on our objective function. From that point, the operator needs to decide how many hotspots λ they are willing to deploy and the trade-off between the monetary cost and the predicted amount of Wi-Fi offloaded traffic.

5.4 PERFORMANCE EVALUATION

To better understand the results of our proposal, hereafter, we have first described the comparison approach then assessed our main results.

5.4.1 *Comparison*

As described in Section 5.2, [105] is the most related work with the one herein proposed. Therefore, we compare our proposal against it which we name "grid-based". In order to create a comparable scenario considering the differences presented on Table 11, we have adapted some aspects of the original scenario presented by the grid-based. It is important to enhance that both the original algorithm and the metric called "data requests" proposed in the grid-based work were applied as originally. The data requests metric ranks the cells based on the number of data requests generated inside the respective cell. Cells in which more data requests were generated are considered to be more capable on providing data offloading and, consequently, are better ranked.

Two are the main changes we have applied to the grid-based scenario in order to create a fair comparison scenario.

First, grid-based traffic model was solely a constant generation of data requests every 5 seconds. The simplicity of this content generation model does not employ any realistic characteristic. Consequently, the offload measurement made on the grid-based work likely does not represent a reliable measurement. Therefore, we have chosen to apply our synthetic traffic model on the grid-based approach in order to have a more precise offload measurement.

Second, grid-based work divides the city on square-shaped cells and restrict one hotspot per cell positioned on its center, resulting on a grid fashion throughout the city. That is not realistic by any means, grid-shaped deployment does not respect the constraints of the city, e. g., a hotspot in the middle of a lake may have its deployment hardly feasible. Therefore, in one of the compared approaches, we have used a subset of our collected PI, limiting one per cell selected as being the one closest to the center of the respective cell. For this purpose, Beijing map was divided into cells of 50 square meters, in a grid shape. To summarize, the approaches we compare against our solution are:

GRID-BASED: The spots selection algorithm, the data request metric and the spot positioning (i. e., center of each cell) are kept as presented in [105]. The synthetic traffic model is the one proposed in Section 5.3.3. Finally, the spots are chosen based on the data requests and their respective offloaded traffic is calculated based on the synthetic traffic.

GRID-BASED (RS): Is the Grid-based with PIs which are a subset of the real collected spots limited to one per cell positioned as close as possible to the cell center. This approach was created in order to use an approximate grid-shaped deployment proposed in [105], while using PIs that are feasible to receive a hotspot.

5.4.2 *Offloaded traffic*

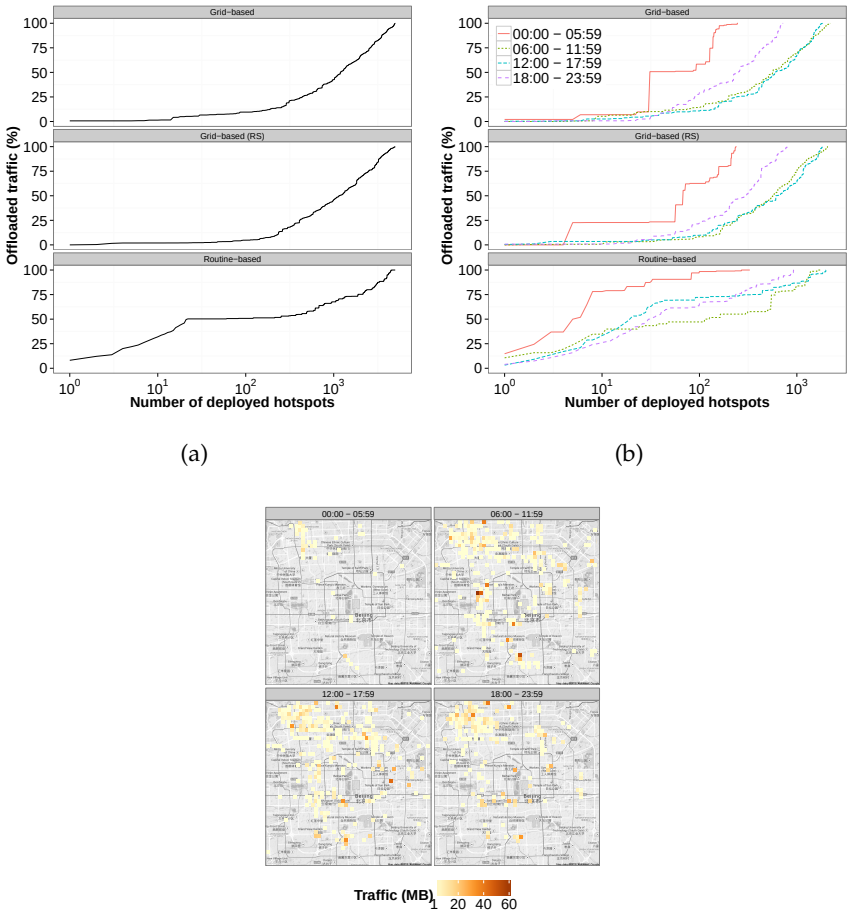
Figures 5.22(a) and 5.22(b) show the offloaded traffic by the different approaches presented on Section 5.4. It is important to note that, to be counted on the graph, a hotspot must have covered at least one person in the simulation scenario. Considering that Grid-based and Grid-based (RS) are limited by one hotspot per cell, their superior limit in number of hotspots is the number of cells in which the hotspots have covered someone. The number of cells which presents both trajectory points and hotspots is 19226. In our evaluation scenario that limitation does not create practical issues for those two approaches because they reach the maximum offload percentage with less than the maximum number of deployed hotspots.

Figure 5.22(a) shows that the routine-based approach has better offload traffic results. For instance, our approach achieves 32%, 49% and 67% of offloading with 10 (< 1%), 50 (1%) and 1000 (20%) deployed hotspots, respectively. For the same amount of hotspots, the Grid-based approach presents 1.3%, 7.3% and 42% of offloaded traffic. This results shows that taking into consideration routine characteristics indeed provides better results than using the data requests approach proposed by the Grid-based. When comparing Grid-based with Grid-based (RS) no significant difference appears. The usage of one real spot per cell positioned on an almost grid-shaped configuration does not increase significantly the efficiency of the Grid-based approach. Indeed, at most there is 4% of difference when comparing those two approaches regarding offloaded data. However, the main difference between them is that it is feasible to deploy a hotspot on a real venue, as considered on the Grid-based (RS), which is not the case on cell centers presented on the Grid-based.

Figure 5.22(b) shows the offloaded traffic for a hotspot deployment configuration that merges result from different periods of the day. Due to the specific characteristics of the mobility on different periods of the day (e. g., period from 00:00 to 05:59 tends to have less trajectories than the period from 12:00 to 17:59), it might be important to assess the deployment per period. Indeed, individual mobile users consume resources in different ways depending on the time and the location that they access the network. Comprehending the differences inherent from different periods of the day allow operators to better allocate network resources.

If it is necessary to guarantee an offload ratio per period, different number of hotspots are needed to offload the traffic on each period of the day. For instance, in order to achieve 80% of offloaded traffic on each period of the day, the routine-based approach employs, respectively, 17 (< 1%), 928 (18.5%), 574 (11.4%) and 308 (6.1%) hotspots for periods from 00:00 to 05:59, from 06:00 to 11:59, from 12:00 to 17:59, and from 18:00 to 23:59. More trajectories likely implies more areas being explored in the city. Consequently, more hotspots are needed to couple with the traffic demands. For the same coverage the grid-based approach needs 136 (2.7%), 1350 (27%), 1348 (26.9%), and 529 (10.5%) hotspots for periods from 00:00 to 05:59, from 06:00 to 11:59, 12:00 to 17:59, and from 18:00 to 23:59, respectively.

Finally, Figure 5.22(c) (better visualized in colors) shows the heatmap of the deployed hotspots using the routine-based approach over Beijing by period of the day. Regardless of the period of the day, a northwest region in the central area consistently shows a concentration of deployed hotspots, which is due to the fact that Microsoft Research Asia headquarter is located there. GeoLife experiment was conduced mostly with Microsoft members that were constantly walking nearby the working place. Therefore, many hotspots were deployed in that region in order to offload the traffic generated there.



(a) (b)

(c)

Figure 22: (Better seen in colors) (a) Offloaded traffic by the number of deployed hotspots per day and (b) period of the day. (c) The heatmap of hotspots by period created with the routine-based approach.

5.5 DISCUSSION

It is important to note that as a first study using real PIs, we do not consider the monetary cost variation per hotspot based on its location. On a completely real scenario the cost might vary depending on the location, because the availability of power and backhaul capacity are widely different at different locations. In this context, it may be economically prohibitive to install hotspots at some of the locations determined to be most useful by the proposed algorithm. On the other hand, it may be very cheap to set up hotspots at other locations, for example, co-located with a cellular base station, fixed-line pay phones, or the homes of DSL customers. However, our solution could be completely applied with an improved version of our database of PIs that would add their respective hotspots' installation cost. Then, our problem would be to find the set of vertices that maximizes the herein proposed metrics with a limited deployment cost ω, i. e., maximize $\sum_{i,t} R_{i,t,m}$ subject to $\sum z_{i,t} < \omega$, where $z_{i,t} = \omega_i$ if i appears on the period t and $z_{i,t} = 0$ if i does not appear in t.

Related to the cost issue, a partially connected topology may present some specific characteristics. For instance, consider a parking area that is not provided with network connection a-priori. The cost of linking the parking area to the operator's network may not pay off the offloaded traffic gains it will produce. A possible solution would be to place other hotspots connected to the network bridging the gap between the parking area and the operator's network. It is important to enhance that our resulting hotspots topology does not guarantee a connected topology, and we think that it would require a prohibitive amount of hotspots to create a completely connected graph covering an urban scenario. However, it may be interesting to assess the creation of connected components in order to provide offload on certain isolated areas.

5.6 CONCLUSIONS

In this chapter, we have presented to the best of our knowledge, the first analysis of a metropolitan-wide hotspot deployment which employs a realistic traffic model. In Section 5.3.1 we have proposed a graph model to represent the relationship between people and the city infra-structure. Furthermore, in Section 5.3.3 we have created a realistic synthetic traffic model to deal with the traffic generation, which takes into consideration transportation modes. Besides, based on common behaviors presented on people's real routines and their traffic demands, we have proposed a methodology to select PIs more suitable to receive hotspot placement data offloading. Our results show that with a small quantity of hotspots it is possible to provide high offload ratio and our routine-based approach provides higher offload ration than the current solution in the literature, e. g., 15% less hotspots and the same offload from 12:00 to 17:59.

6

CONCLUSIONS AND FUTURE HORIZONS

"The future cannot be predicted, but futures can be invented."

— Dennis Gabor

6.1 SUMMARY

Predictions to the mobile data traffic growth are unison: it will continue in the years to come. In this thesis we argue that the key to cope with this trend is to understand both users' network traffic and mobility behavior. In order to understand the users we have focused on a natural characteristic of people: routine.

Literature presented some scattered insights with regards to the cyclicity present on human mobility and network traffic. We extensively evaluated both topics in the context of an urban scenario. We have started by creating a common representation of fine- and coarse-grained trajectories and their interactions with the urban environment. Next, we characterized human mobility from their visit, temporal, and spatiotemporal behavior in several cities using mostly fine-grained mobility datasets. This lead us to comprehend three core aspects in the urban human mobility. First, there is a repetitive pattern in the visiting behavior concentrated to few locations. Repetitiveness metric showed that few percentage of locations (points of interest or cell tower antennas) receive high percentage of the visits. Second, people's displacement is generally short, three-quarters of trajectories in all studied cities are up to 10 km. Finally, there is a strong tendency to use the shortest-paths, the majority of the trajectories are only half longer than the possible shortest-path.

Aside from that, several specificities have been unveiled in people's mobility, e. g., more time spent when visiting venues and larger displacement as the week passes by, more places visited during weekends than weekdays, and less displacement during the mornings of weekends when compared to weekdays due to the more active Friday and Saturday nights. Besides, we have shown that human mobility does not respond only to a rigid set of rules regardless the city context. For instance, contrary to Beijing, in Paris, people spent 66% less time in shopping-related venues during weekends as compared to weekdays, due to a limited number of stores opened during the whole weekend.

The second focus of this thesis is the characterization, classification, and synthetic generation of subscribers using their network traffic demands. From the characterization we have seen daily cyclicity

for users' traffic measurements. Therefore, we have selected and further studied one typical day from the dataset. We have classified subscribers by their traffic volume, number of session, and the session inter-arrival time similarity. The classification resulted in six subscribers' network usage behavior profiles. Since each of the profiles presented considerable network usage differences in certain hours of the day, we have further separated and characterized their network activity in peak and non-peak hours. Finally, we have synthetically reproduced the network traffic demands of those six different profiles of subscribers in two time periods of a routinary typical day.

Besides, network traffic characterization brought interesting outcomes. For instance, session volumes generally appear in three ways: upload and download are very similar, the former is high and the latter is almost non-existent, or vice-versa. Additionally, in the studied week, two groups of subscribers summed up together represent the majority, the ones that generate traffic everyday and the ones that generate traffic in only one of the days. Finally, in the studied dataset, subscribers' age and gender influence their network activity in two main ways: older ones generate less traffic demands than younger ones, and male activity is higher than the female counterparts.

6.2 FUTURE HORIZONS

Looking ahead, we see a wide range of possible research direction in both short- and long-term bond to the human mobility and network traffic routine analysis. Short-term directions relate to improvements in the models and analysis present in this thesis. Long-term refer mostly to next steps that can take advantage of insights proposed in this thesis.

6.2.1 *Short-term*

MOBILITY FROM DIFFERENT NATURE: Even if we have presented mobility analysis from both CDR- and GPS-based trajectories, we envision other possibilities. Services such as Foursquare and Instagram provide large-scale human data collection, which often contains geolocalized information. Similar to a CDR record, user positioning is only available when he performs an activity, which in this context indicates his presence (e. g., *check-in* on Foursquare) at a certain location. This comprises a whole area of research called Participatory Sensing Networks [125, 126]. As with a CDR dataset, a sequence of check-ins represent a user trajectory, e. g., simply temporally concatenating them, or by inference [127]. The human mobility and routine analysis from this source is still unexplored, to the best of our knowledge.

TRAFFIC MODEL: We aim to investigate models to describe sessions' transfer rate and duration. Although not crucial parameters, they can enrich the model and the understanding of usage profiles. Besides, better knowledge about transfer rate may help the operator, among other actions, to better allocate resources, improve Quality of Service, and to provide diversify data plans. Session duration may help to identify anomalous behavior or to identify disconnection issues.

STUDY CASE: As with any simulation, our study case abstracts some characteristics of a real world implementation. For instance, adding a fading model to our coverage definition may improve the precision of the coverage time, visiting count, and offload ratio metrics. Besides, as a first evaluation of a real-world scenario hotspot deployment, we have neglected the financial cost variation imposed by the deployment in different areas, e. g., venues already connected by the operator's network (or landline) have supposedly lower deployment cost compared to a non-connected venue. Such problem has multi-objectives, in which the model tries to minimize the financial cost and maximize the benefit (the amount of offloaded data in our problem). In order to perform such evaluation, the dataset must contain the deployment cost for venues that are connected and non-connected by the operator's network. Such information may be unfeasible to obtain from Telcos, thus we can evaluate simple scenarios, e. g., different percentages of non-connected and connected venues that are randomly chosen among the PIs. By varying the cost ratio between connected and non-connected we obtain an estimation of the total cost-benefit of the deployment. Still, it is a challenging task to model the cost difference between the deployment of a hotspot on a connected and on a non-connected point of interest.

Besides, we envision to evaluate the performance of the hotspot deployment topology created with our routine-based strategy using the mobility from a period which is later in time. For example, our results were made using a topology that took into consideration mobility data from November to December 2008, how good is the performance if we apply the same topology for the mobility taken one year later? As the participants of the GeoLife experiment change during time, i. e., some left and some joined the experiment, we are probable going to see some deterioration on the offload performance. It is due to the possible mobility changes, which the first topology did not took into consideration. The outcomes depend on how variable is the set of people being evaluated. For short periods of time, however, we foresee little change and consequently, almost unchanged performance if considering a large scenario with millions of users.

One possible solution is to select new venues and deploy additional hotspots in the circumstances of more drastic mobility changes, which were not existent in the first deployment.

6.2.2 *Long-term*

DATASET COLLECTION: As a vast spectrum of work can take advantage of dataset analysis in the area of human mobility and network, experiments aiming to collect rich information from users are of enormous value for the research community. Information regarding user, his device, and surrounding scenario, e. g., fine-grained mobility, battery level, running applications, access points, bluetooth devices, network traffic usage can contribute to a deeper understanding on how we interact with the network and with the environment around. For instance, except for MACACO [128] and Priva'Mov [129] projects, to the best of our knowledge, no other experiment makes effort in this direction. Besides, in order to further provide positive impact, such projects may increase the number of participating entities, thus to broaden the research fields that can take advantage of the collected data.

FORWARDING PROTOCOLS: An important problem on intermittently connected networks is how to couple human mobility patterns with message forwarding algorithms [21]. In this area, mobility has been widely studied when it comes down to encounters among nodes. Contrarily, trajectories behavior could be studied to improve protocols based on store-and-forward late delivery. How far a person routinely goes or how frequently one visits the same places in a city it is an important characteristics to define the potential of a peer to be chosen to keep a message to be routed. For instance, important routers on a pocket switched network could be taken from the set of few individuals routinely going further than the 10 km a day, thus carrying the messages farther.

ROUTINE-BASED DATA PREFETCHING: Prefetching has been shown to effectively reduce user perceived latency. An interesting approach might be to merge trajectory information with demanded content in order to provide a smart data prefetching approach. Based on the routine of a user, the prefetching service could store static content of often visited web pages. This content could be downloaded using access points present in the locations he routinely visits. For instance, an application for mobile phones could download content in advance at home using WiFi and this would be later presented along with content downloaded using the mobile cellular network in real time. This ap-

proach could alleviate the usage of the cellular network by shifting to inexpensive networks (such as WiFi) the load of static content such as images.

URBAN PLANNING: Monitoring, distributing, and processing traffic information may enable better strategic planning and encourage better use of public transportation. Several applications may take advantage of the driver's routinary behavior in order to improve aspects of the vehicular networks. For instance, traffic information, e. g., accidents, construction sites, traffic jams may be exchanged between vehicles. On a routinary scenario, the human mobility characteristics considered in this thesis can be used to forecast the situation for the next days and inform the driver, for example, about possible alternative roads. Besides, a service may identify points of interest based on driver's mobility patterns. Identify parking lots and its availability in number of free spots, or suggest the best charging station for electrical vehicles based on the driver's routine and battery conditions. On a collaborative scenario, a carpooling service could suggest people to get or to offer a ride based on regular driver's destinations and passengers interested on ride-sharing.

CUSTOMIZED AD-CAMPAIGN: A service to advertise products on roadside signs may merge different sources of information in order to display targeted marketing. By crossing information from the people's routinary trajectories, interests, and traffic conditions, a service could present ads that match people's interest on a certain area of the city likely having him on its vicinity.

APPENDIX

A.1 CLASSES AND CATEGORIES FOR POINTS OF INTEREST

Table 13: Classes and their respective categories

Class	Abbr.	Categories
Arts & Entertainment	A/E	amusement park, aquarium, art gallery, bowling alley, casino, movie rental, movie theater, museum, zoo
Education	Edu	school, university, library
Food	Food	bakery, cafe, food, grocery, meal delivery, meal takeaway, restaurant
Religion	Rel	church, hindu temple, mosque, place of worship, synagogue
Outdoor & Sports	O/S	campground, cemetery, gym, park, stadium
Night Life	NL	bar and night club
Shopping	Shop	bicycle store, book store, car dealer, clothing store, convenience store, department store, electronics store, establishment, florist, furniture store, hardware store, home goods store, jewelry store, liquor store, market, pet store, pharmacy, shoe store, shopping mall, store, supermarket
Travel	Trvl	airport, bus station, embassy, lodging, parking, rv park, subway station, taxi stand, train station, travel agency
Services	Srvc	accounting, atm, bank, beauty salon, car rental, car repair, car wash, dentist, doctor, electrician, funeral home, gas station, hair care, health, hospital, insurance agency, laundry, lawyer, locksmith, painter, physiotherapist, plumber, police, post office, real estate agency, roofing contractor, spa, storage, city hall, courthouse, finance, fire station, moving company, general contractor, veterinary care, local government office

A.2 CDFS OF THE TRAFFIC PARAMETERS IN PEAK AND NON-PEAK HOURS

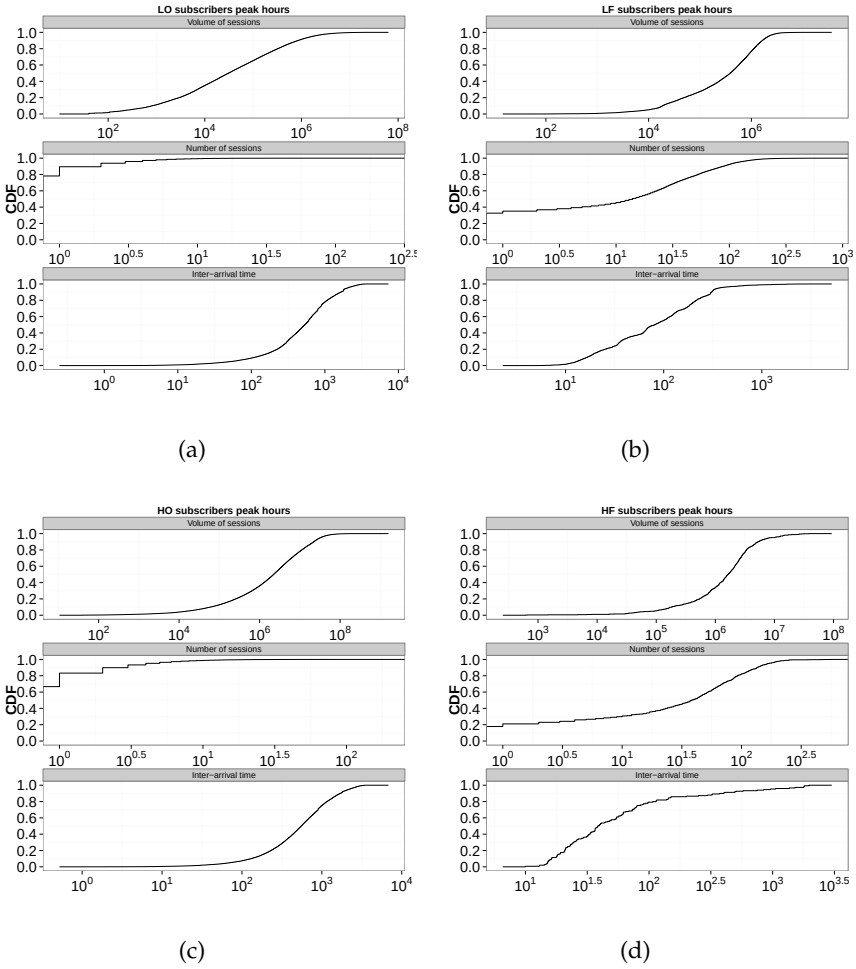


Figure 23: (a) Traffic parameters for LO users in peak hours. (b) Traffic parameters for LF users in peak hours. (c) Traffic parameters for HO users in peak hours. (d) Traffic parameters for HF users in peak hours.

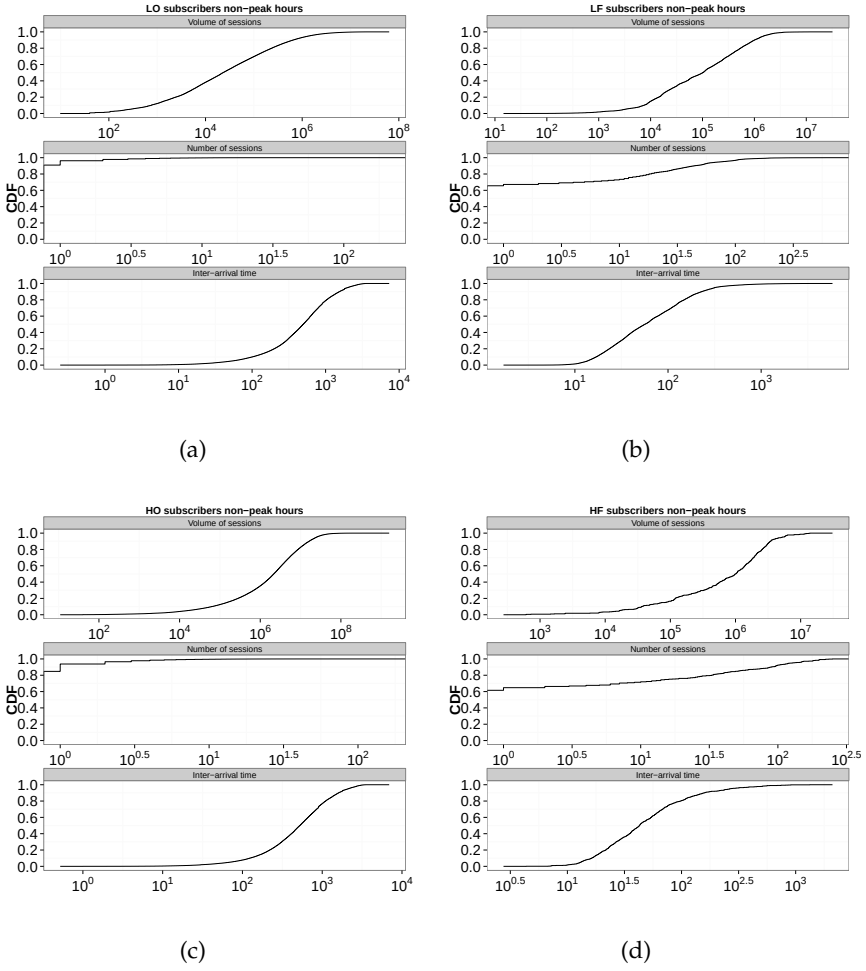


Figure 24: (a) Traffic parameters for LO users in non-peak hours. (b) Traffic parameters for LF users in non-peak hours. (c) Traffic parameters for HO users in non-peak hours. (d) Traffic parameters for HF users in non-peak hours.

A.3 SYNTHETIC TRAFFIC GENERATOR ALGORITHM

Algorithm 1 Synthetic traffic generator algorithm

```
1:  procedure GENERATE(n)
2:      for each integer i from 1 to n do
3:          class_i = CLASS()
4:          for t in D do
5:              for Metric in {N, IAT, V} do
6:                  pMetric_i^t = DISTR(Metric, t, class_i)
7:              end for
8:              ▷ Get the number of sessions
9:              nrsess = GET(pN_i^t, 1)
10:             Arrivals_i = [ ]
11:             ▷ Get an array of IATs with length = 'number of
                  session', i. e., one arrival per session
12:             AvgIAT = GET(pIAT_i^t, nrsess)
13:             for intarr in AvgIAT do
14:                 if intarr == AvgIAT[0] then
15:                     arr = intarr
16:                 else
17:                     arr = AvgIAT[intarr − 1] + intarr
18:                 end if
19:                 APPEND(Arrivals, arr)
20:             end for
21:             ▷ Get an array of Volumes with length = 'num-
                  ber of session', i. e., one volume per session
22:             Volumes_i = GET(pIAT_i^t, nrsess)
23:         end for
24:         Sessions_i = [ ]
25:         for arrival, volume in Arrivals_i, Volumes_i do
26:             session = SESSION(arrival, volume)
27:             APPEND(Sessions_i, session)
28:         end for
          return Sessions_i
29:     end for
30: end procedure
```

BIBLIOGRAPHY

[1] Mark Weiser. "The Computer for the 21st Century." In: *Scientific American* 265.3 (Jan. 1991), pp. 66–75.

[2] Dave Evans. *The Internet of Things How the Next Evolution of the Internet Is Changing Everything*. CISCO White Paper. Apr. 2011.

[3] Janessa Rivera and Rob van der Meulen. *Gartner Says Annual Smartphone Sales Surpassed Sales of Feature Phones for the First Time in 2013*. English. Gartner. Feb. 2013. URL: http://www.gartner.com/newsroom/id/2665715.

[4] Cisco. *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2014–2019*. Feb. 2014.

[5] Instagram. *Instagram statistics*. English. Instagram. 2014. URL: http://instagram.com/press/.

[6] Facebook, Ericsson, and Qualcomm. *A Focus on Efficiency*. English. Internet.org. Sept. 2013.

[7] WhatsApp. *WhatsApp statistics*. English. WhatsApp. Apr. 2014. URL: http://blog.whatsapp.com/613/500000000.

[8] Twitter. *Twitter statistics*. English. Twitter. 2014. URL: https://blog.twitter.com/2014/the-2014-yearontwitter.

[9] Youtube. *YouTube statistics*. English. YouTube. 2014. URL: http://youtube.com/yt/press/statistics.html.

[10] John Donovan. *AT&T's data traffic is actually doubling annually*. English. AT&T. Feb. 2012.

[11] Marc Sier, Stephan Kalleder, and Andreas Pauly. *Mobile Data Growth: How Operators Can Handle The Traffic Explosion*. English. Solon. Apr. 2012.

[12] GSMA Intelligence. *The Mobile Economy 2014*. Report.

[13] Kyunghan Lee et al. "Mobile Data Offloading: How Much Can WiFi Deliver?" In: *Networking, IEEE/ACM Transactions on* 21.2 (Apr. 2013), pp. 536–550.

[14] K. Thilakarathna, H. Petander, and A. Seneviratne. "Performance of content replication in MobiTribe: A distributed architecture for mobile UGC sharing." In: *Local Computer Networks (LCN), 2011 IEEE 36th Conference on*. Oct. 2011, pp. 554–562.

[15] Annalisa Socievole, Floriano De Rango, and Antonio Caputo. "Wireless contacts, Facebook friendships and interests: Analysis of a multi-layer social network in an academic environment." In: *Wireless Days (WD), 2014 IFIP*. Nov. 2014, pp. 1–7.

[16] Long Vu et al. "Joint Bluetooth/Wifi Scanning Framework for Characterizing and Leveraging People Movement in University Campus." In: *Proceedings of the 13th ACM International Conference on Modeling, Analysis, and Simulation of Wireless and Mobile Systems*. MSWIM '10. Bodrum, Turkey: ACM, 2010, pp. 257–265.

[17] Augustin Chaintreau et al. *Pocket Switched Networks: Real-world mobility and its consequences for opportunistic forwarding*. Tech. rep. UCAM-CL-TR-617. University of Cambridge, Computer Laboratory, Feb. 2005.

[18] Nathan Eagle and Alex S. Pentland. "Reality mining: sensing complex social systems." In: *Personal Ubiquitous Comput.* 10.4 (May 2006), pp. 255–268.

[19] Yohan Chon and Hojung Cha. "LifeMap: A Smartphone-Based Context Provider for Location-Based Services." In: *Pervasive Computing, IEEE* 10.2 (Apr. 2011), pp. 58–67.

[20] Anna-Kaisa Pietiläinen and Christophe Diot. "Dissemination in Opportunistic Social Networks: The Role of Temporal Communities." In: *Proceedings of the Thirteenth ACM International Symposium on Mobile Ad Hoc Networking and Computing*. MobiHoc '12. Hilton Head, South Carolina, USA: ACM, 2012, pp. 165–174.

[21] Pan Hui et al. "Pocket Switched Networks and Human Mobility in Conference Environments." In: *Proceedings of the 2005 ACM SIGCOMM Workshop on Delay-tolerant Networking*. WDTN '05. Philadelphia, Pennsylvania, USA: ACM, 2005, pp. 244–251.

[22] Adriano Galati, Karim Djemame, and Chris Greenhalgh. "A mobility model for shopping mall environments founded on real traces." English. In: *Networking Science* 2.1-2 (2013), pp. 1–11.

[23] N. Kiukkonen et al. "Towards rich mobile phone datasets: Lausanne data collection campaign." In: *Proc. ACM Int. Conf. on Pervasive Services*. July 2010.

[24] Yu Zheng, Xing Xie, and Wei-Ying Ma. "GeoLife: A Collaborative Social Networking Service among User, Location and Trajectory." In: *IEEE Data Eng. Bull.* 33.2 (2010), pp. 32–39.

[25] Michal Piorkowski, Natasa Sarafijanovic-Djukic, and Matthias Grossglauser. "A parsimonious model of mobile partitioned networks with clustering." In: *Communication Systems and Networks and Workshops, 2009. COMSNETS 2009. First International*. Bangalore, India: IEEE, Jan. 2009, pp. 1–10.

[26] Raul Amici et al. “Performance Assessment of an Epidemic Protocol in {VANET} Using Real Traces.” In: *Procedia Computer Science* 40 (2014). Fourth International Conference on Selected Topics in Mobile & Wireless Networking (MoWNet 2014), pp. 92–99.

[27] Ratul Mahajan, John Zahorjan, and Brian Zill. “Understanding Wifi-based Connectivity from Moving Vehicles.” In: *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*. IMC ’07. San Diego, CA, USA: ACM, 2007, pp. 321–326.

[28] Jason LeBrun and Chen N. Chuah. “Bluetooth content distribution stations on public transit.” In: *Proceedings of the 1st international workshop on Decentralized resource sharing in mobile computing and networking - MobiShare ’06*. MobiShare ’06. Los Angeles, CA, USA: ACM Press, Sept. 2006, pp. 63–65.

[29] J. Jetcheva et al. “Design and evaluation of a metropolitan area multitier wireless ad hoc network architecture.” In: *Fifth IEEE Workshop on Mobile Computing Systems and Applications*. Monterey, CA, USA: IEEE, 2003, pp. 32–43.

[30] Huayong Wang et al. “Transportation mode inference from anonymized and aggregated mobile phone call detail records.” In: *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*. Sept. 2010, pp. 318–323.

[31] Chaoming Song et al. “Limits of Predictability in Human Mobility.” In: *Science* 327.5968 (Feb. 19, 2010), pp. 1018–1021.

[32] Albert-Laszlo Barabasi. “The origin of bursts and heavy tails in human dynamics.” In: *Nature* 435 (2005), p. 207.

[33] Gyan Ranjan et al. “Are Call Detail Records Biased for Sampling Human Mobility?” In: *SIGMOBILE Mob. Comput. Commun. Rev.* 16.3 (Dec. 2012), pp. 33–44.

[34] F. Ben Abdesslem and A. Lindgren. “Large scale characterisation of YouTube requests in a cellular network.” In: *A World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2014 IEEE 15th International Symposium on*. June 2014, pp. 1–9.

[35] Gabriel Ramos-Fernández et al. “Lévy walk patterns in the foraging movements of spider monkeys (Ateles geoffroyi).” In: *Behavioral Ecology and Sociobiology* 55.3 (Jan. 1, 2004), pp. 223–230.

[36] R. P. D. Atkinson et al. “Scale-Free Dynamics in the Movement Patterns of Jackals.” English. In: *Oikos* 98.1 (2002),

[37] G. M. Viswanathan et al. “Lévy flight search patterns of wandering albatrosses.” In: *Nature* 381.6581 (May 30, 1996), pp. 413–415.

[38] M. F. Shlesinger, J. Klafter, and Gert Zumofen. "Above, below and beyond Brownian motion." In: *Am. J. Phys.* 67.12 (Dec. 1999).

[39] D Brockmann, L Hufnagel, and T Geisel. "The scaling laws of human travel." In: *Nature* 439.7075 (Jan. 2006), pp. 462–465.

[40] I. Rhee et al. *On the Levy-walk nature of human mobility: Do humans walk like monkeys?* 2007.

[41] Marta C. Gonzalez, Cesar A. Hidalgo, and Albert-Laszlo Barabasi. "Understanding individual human mobility patterns." In: *Nature* 453.7196 (June 2008), pp. 779–782.

[42] Chaoming Song et al. "Modelling the scaling properties of human mobility." In: *Nature Physics* 6.10 (Sept. 12, 2010), pp. 818–823.

[43] Sibren Isaacman et al. "Identifying Important Places in People's Lives from Cellular Network Data." In: *Proceedings of the 9th International Conference on Pervasive Computing*. Pervasive'11. San Francisco, USA: Springer-Verlag, 2011, pp. 133–151.

[44] R.A Becker et al. "A Tale of One City: Using Cellular Network Data for Urban Planning." In: *IEEE Pervasive Computing* 10.4 (Apr. 2011), pp. 18–26.

[45] Yunji Liang et al. "Understanding the Regularity and Variability of Human Mobility from Geo-trajectory." In: *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2012 IEEE/WIC/ACM International Conferences on*. Vol. 1. Dec. 2012, pp. 409–414.

[46] S. Motahari, Hui Zang, and P. Reuther. "The Impact of Temporal Factors on Mobility Patterns." In: *System Science (HICSS), 2012 45th Hawaii International Conference on*. Jan. 2012, pp. 5659–5668.

[47] Vania Conan, Jérémie Leguay, and Timur Friedman. "Characterizing Pairwise Inter-contact Patterns in Delay Tolerant Networks." In: *Proceedings of the 1st International Conference on Autonomic Computing and Communication Systems*. Autonomics '07. Rome, Italy: ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2007, 19:1–19:9.

[48] A. Chaintreau et al. "Impact of Human Mobility on Opportunistic Forwarding Algorithms." In: *Mobile Computing, IEEE Transactions on* 6.6 (June 2007), pp. 606–620.

[49] Anirudh Natarajan, Mehul Motani, and Vikram Srinivasan. "Understanding Urban Interactions from Bluetooth Phone Contact Traces." In: *Proceedings of the 8th International Conference on Passive and Active Network Measurement*. PAM'07. Louvain-la-Neuve, Belgium: Springer-Verlag, 2007, pp. 115–124.

[50] Yi Wang, B. Krishnamachari, and T.W. Valente. "Findings from an empirical study of fine-grained human social contacts." In: *Wireless On-Demand Network Systems and Services, 2009. WONS 2009. Sixth International Conference on*. Feb. 2009, pp. 153–160.

[51] Thomas Kunz et al. "WAP Traffic: Description and Comparison to WWW Traffic." In: *Proceedings of the 3rd ACM International Workshop on Modeling, Analysis and Simulation of Wireless and Mobile Systems*. MSWIM '00. Boston, Massachusetts, USA: ACM, 2000, pp. 11–19.

[52] Atul Adya, Paramvir Bahl, and Lili Qiu. "Analyzing the browse patterns of mobile clients." In: *Internet Measurement Workshop*. Ed. by Vern Paxson. ACM, Feb. 24, 2005, pp. 189–194.

[53] C. Williamson et al. "Characterization of CDMA2000 cellular data network traffic." In: *Local Computer Networks, 2005. 30th Anniversary. The IEEE Conference on*. Nov. 2005, Z000–719.

[54] J. Candia et al. "Uncovering individual and collective human dynamics from mobile phone records." In: *Journal of Physics A: Mathematical and Theoretical* 41 (2008).

[55] Diane Tang and Mary Baker. "Analysis of a Metropolitan-area Wireless Network." In: *Proceedings of the 5th Annual ACM/IEEE International Conference on Mobile Computing and Networking*. MobiCom '99. Seattle, Washington, USA: ACM, 1999, pp. 13–23.

[56] C. W. Omlin O. A. Abidogun. "A self organizing maps model for outlier detection in call data from mobile telecommunication networks." In: *Proc. of the Southern Africa Telecommunication Networks and Applications Conference (SATNAC)*. Aug. 2004.

[57] Alec Pawling, Nitesh V. Chawla, and Greg Madey. "Anomaly detection in a mobile communication network." In: *Computational and Mathematical Organization Theory* 13.4 (2007), pp. 407–422.

[58] P.A Vijaya, M Narasimha Murty, and D.K Subramanian. "Leaders-Subleaders: An efficient hierarchical clustering algorithm for large data sets." In: *Pattern Recognition Letters* 25.4 (2004), pp. 505–513.

[59] S. Lloyd. "Least squares quantization in PCM." In: *Information Theory, IEEE Transactions on* 28.2 (Mar. 1982), pp. 129–137.

[60] Richard A. Becker et al. *Clustering Anonymized Mobile Call Detail Records to Find Usage Groups*. Workshop on Pervasive and Urban Applications (PURBA). 2011.

[61] Amanda Lenhart et al. *Teens and Mobile Phones*. Tech. rep. Pew Research Center, Apr. 2010.

[62] D. Naboulsi, R. Stanica, and M. Fiore. "Classifying call profiles in large-scale mobile traffic datasets." In: *Proc. of IEEE Infocom*. Apr. 2014.

[63] Ram Keralapura et al. "Profiling Users in a 3G Network Using Hourglass Co-clustering." In: *Proc. of ACM MobiCom*. Sept. 2010.

[64] Sahar Hoteit et al. "Content Consumption Cartography of the Paris Urban Region Using Cellular Probe Data." In: *Proc. of the 1st Workshop on Urban Networking (ACM UrbaNe)*. Dec. 2012.

[65] P. Paraskevopoulos et al. "Identification and Characterization of Human Behavior Patterns from Mobile Phone Data." In: *Proc. of NetMob*. May 2013.

[66] F. Girardin et al. "Towards estimating the presence of visitors from the aggragate mobile phone network activity they generate." In: *Proc. of Intl. Conference on Computers in Urban Planning and Urban Management*. 2009.

[67] R. M. Pulselli et al. "Computing urban mobile landscapes through monitoring population density based on cellphone chatting." In: *Int. Journal of Design and Nature and Ecodynamics* 3 (2008).

[68] A Vaccari et al. "A holistic framework for the study of urban traces and the profiling of urban processes and dynamics." In: *Proc. of Int. IEEE Conf. on Intelligent Transportation Systems (ITSC)*. Oct. 2009.

[69] Daniel S. Hamermesh. *Routine*. Working Paper 9440. National Bureau of Economic Research, Jan. 2003.

[70] James A. Throgmorton and Barbara. Eckstein. *Desire Lines: The Chicago Area Transportation Study and the Paradox of Self in Post-War America*. Nov. 2000. URL: http://www.nottingham.ac.uk/3cities/throgeck.htm.

[71] Zhong-Ren Peng. "Urban transportation strategies in Chinese cities and their impacts on the urban poor." In: *Transportation Research Board 85th Annual Meeting* (2005), p. 14.

[72] Jia Lin. "Bicycles in Beijing." Project: Transportation (Green Design and the City).

[73] Shuling Wang et al. "Evaluation on Vehicle Restriction Measure in Beijing." In: *Traffic and Transportation Studies 2010*. Chap. 39, pp. 433–443.

[74] Zheng Xin. *Subway Line 6 to start running in December*. English. China Daily. Nov. 2012. URL: http://www.chinadaily.com.cn/beijing/2012-11/26/content_15998073.htm.

[75] ISO. *ISO 8601:1988. Data elements and interchange formats — Information interchange — Representation of dates and times*. See also 1-page correction, ISO 8601:1988/Cor 1:1991. 1988, p. 14.

[76] D. B. Carr, A. R. Olsen, and D. White. "Hexagon mosaic maps for displaying univariate and bivariate geographical data." In: *Cartography & Geographical Information Systems* 19 (1992), pp. 228–236.

[77] P. A. P. Moran. "Notes on Continuous Stochastic Phenomena." In: *Biometrika* 37.1/2 (June 1950), pp. 17–23.

[78] Eduardo Mucelli Rezende Oliveira and Aline Carneiro Viana. "From Routine to Network Deployment for Data Offloading in Metropolitan Areas." In: *Proc. of IEEE SECON*. June 2014.

[79] Jenna Wortham. "Cellphones Now Used More for Data Than for Calls." In: *New York Times* (May 2010).

[80] Alcatel-Lucent. *Alcatel-Lucent 9900 Wireless Network Guardian*. White Paper. Dec. 2012.

[81] U. Paul et al. "Understanding traffic dynamics in cellular data networks." In: *Proc. of IEEE Infocom*. Apr. 2011.

[82] International Trade Union Confederation. *Frozen in time: Gender pay gap unchanged for 10 years*. Tech. rep. 2012.

[83] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. "On Clustering Validation Techniques." In: *Journal of Intelligent Information Systems* 17.2-3 (Dec. 2001), pp. 107–145.

[84] R. R. Sokal and C. D. Michener. "A statistical method for evaluating systematic relationships." In: *University of Kansas Scientific Bulletin* 28 (1958), pp. 1409–1438.

[85] GlennW. Milligan and MarthaC. Cooper. "An examination of procedures for determining the number of clusters in a data set." In: *Psychometrika* 50.2 (1985), pp. 159–179.

[86] J. C. Dunn. "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters." In: *Journal of Cybernetics* 3.3 (1973), pp. 32–57.

[87] W. J. Krzanowski and Y. T. Lai. "A Criterion for Determining the Number of Groups in a Data Set Using Sum-of-Squares Clustering." In: *Biometrics* 44.1 (Mar. 1988), pp. 22–34.

[88] Maria Halkidi, Michalis Vazirgiannis, and Yannis Batistakis. "Quality Scheme Assessment in the Clustering Process." In: *Proc. of the 4th European Conf. on Principles of Data Mining and Knowledge Discovery*. Sept. 2000.

[89] Peter Rousseeuw. "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis." In: *Elsevier Journal of Computational Applied Mathematics* 20.1 (Nov. 1987), pp. 53–65.

[90] T. Hossmann, T. Spyropoulos, and F. Legendre. "Know Thy Neighbor: Towards Optimal Mapping of Contacts to Social Graphs for DTN Routing." In: *INFOCOM, 2010 Proceedings IEEE*. Mar. 2010, pp. 1–9.

[91] Ralph B. D'Agostino and Michael A. Stephens. *Goodness-of-Fit-Techniques*. Vol. 68. CRC Press, June 1986.

[92] Karl Pearson. "X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling." In: *Philosophical Magazine Series 5* 50.302 (1900), pp. 157–175.

[93] A. Bhattacharyya. "On a measure of divergence between two statistical populations defined by their probability distributions." In: *Bulletin of the Calcutta Mathematical Society* 35 (1943), pp. 99–109.

[94] D. Comaniciu, V. Ramesh, and P. Meer. "Kernel-based object tracking." In: *IEEE Trans. on Pattern Analysis and Machine Intelligence* 25.5 (May 2003), pp. 564–577.

[95] S. Kullback and R. A. Leibler. "On information and sufficiency." In: *Annals of Mathematical Statistics* 22 (1951), pp. 49–86.

[96] President's Council of Advisors on Science and Technology. *Big Data and Privacy: A Technological Perspective*. Tech. rep. Executive Office of the President, May 2014.

[97] Gartner. *Gartner Says Worldwide Smartphone Sales Soared in Fourth Quarter of 2011 With 47 Percent Growth*. Gartner. Apr. 2011. URL: http://www.gartner.com/it/page.jsp?id=1924314.

[98] Cisco. *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2011–2017*. Cisco. Feb. 2012. URL: http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.html.

[99] AT&T. *AT&T Launches Pilot Wi-Fi Project in Times Square*. AT&T. May 2010. URL: http://www.att.com/gen/press-room?pid=17961&cdvn=news&newsarticleid=30838.

[100] Jenna Wortham. *Customers Angered as iPhones Overload AT&T*. The New York Times. Sept. 2009. URL: http://www.nytimes.com/2009/09/03/technology/companies/03att.html.

[101] Karen C. Seto et al. "A Meta-Analysis of Global Urban Land Expansion." In: *PLoS ONE* 6.8 (Aug. 2011), e23777.

[102] Glenn Judd, Xiaohui Wang, and Peter Steenkiste. "Efficient channel-aware rate adaptation in dynamic environments." In: *Proceedings of the 6th international conference on Mobile systems, applications, and services*. MobiSys '08. Breckenridge, CO, USA: ACM, 2008, pp. 118–131.

[103] M. Kim, D. Kotz, and S. Kim. "Extracting a Mobility Model from Real User Traces." In: *INFOCOM 2006. 25th IEEE International Conference on Computer Communications. Proceedings*. Apr. 2006, pp. 1–13.

[104] Tian Wang et al. "Efficient WiFi deployment algorithms based on realistic mobility characteristics." In: *IEEE MASS*. Nov. 2010, pp. 422–431.

[105] Eyuphan Bulut and Boleslaw K. Szymanski. "WiFi Access Point Deployment for Efficient Mobile Data Offloading." In: *Proceedings of the First ACM International Workshop on Practical Issues and Applications in Next Generation Wireless Networks*. PINGEN '12. Istanbul, Turkey: ACM, 2012, pp. 45–50.

[106] H. Klessig, M. Gunzel, and G. Fettweis. "Increasing the Capacity of Large-Scale HetNets through Centralized Dynamic Data Offloading." In: *Vehicular Technology Conference (VTC Fall), 2014 IEEE 8oth*. Sept. 2014, pp. 1–7.

[107] Liang Hu et al. "Realistic Indoor Wi-Fi and Femto Deployment Study as the Offloading Solution to LTE Macro Networks." In: *Vehicular Technology Conference (VTC Fall), 2012 IEEE*. Sept. 2012, pp. 1–6.

[108] A. Furuskar, M. Almgren, and K. Johansson. "An infrastructure cost evaluation of single- and multi-access networks with heterogeneous traffic density." In: *Vehicular Technology Conference, 2005. VTC 2005-Spring. 2005 IEEE 61st*. Vol. 5. May 2005, pp. 3166–3170.

[109] K. Johansson and A. Furuskar. "Cost efficient capacity expansion strategies using multi-access networks." In: *Vehicular Technology Conference, 2005. VTC 2005-Spring. 2005 IEEE 61st*. Vol. 5. May 2005, pp. 2989–2993.

[110] K. Johansson, J. Zander, and A. Furuskar. "Cost Efficient Deployment of Heterogeneous Wireless Access Networks." In: *Vehicular Technology Conference, 2007. VTC2007-Spring. IEEE 65th*. Apr. 2007, pp. 3200–3204.

[111] Tristan Henderson, David Kotz, and Ilya Abyzov. "The changing usage of a mature campus-wide wireless network." In: *Computer Networks* 52 (14 Oct. 2008), pp. 2690–2712.

[112] A. Kchiche and F. Kamoun. "Access-Points Deployment for Vehicular Networks Based on Group Centrality." In: *New Technologies, Mobility and Security (NTMS), 2009 3rd International Conference on*. 2009, pp. 1–6.

[113] A. Kchiche and F. Kamoun. "Centrality-based Access-Points deployment for vehicular networks." In: *Telecommunications (ICT), 2010 IEEE 17th International Conference on*. 2010, pp. 700–706.

[114] M.Z. Chowdhury et al. "Service quality improvement of mobile users in vehicular environment by mobile femtocell network deployment." In: *ICT Convergence (ICTC), 2011 International Conference on*. 2011, pp. 194–198.

[115] S. Dimatteo et al. "Cellular Traffic Offloading through WiFi Networks." In: *Mobile Adhoc and Sensor Systems (MASS), 2011 IEEE 8th International Conference on*. 2011, pp. 192–201.

[116] M. V. Barbera et al. "VIP delegation: Enabling VIPs to offload data in wireless social mobile networks." In: *Distributed Computing in Sensor Systems and Workshops (DCOSS), 2011 International Conference on*. IEEE, June 2011, pp. 1–8.

[117] Bo Han et al. "Mobile Data Offloading through Opportunistic Communications and Social Participation." In: *Mobile Computing, IEEE Transactions on* 11.5 (May 2012), pp. 821–834.

[118] F.Rebecchi et al. "Data Offloading Techiniques in Cellular Networks: A Survey." In: *IEEE Communications Surveys and Tutorials* 99 (June 2014), pp. 1–25.

[119] John Whitbeck et al. "Fast Track Article: Push-and-track: Saving Infrastructure Bandwidth Through Opportunistic Forwarding." In: *Pervasive Mob. Comput.* 8.5 (Oct. 2012), pp. 682–697.

[120] Alfonso Shimbel. "Structural parameters of communication networks." English. In: *The bulletin of mathematical biophysics* 15.4 (1953), pp. 501–507.

[121] Gert Sabidussi. "The centrality index of a graph." English. In: *Psychometrika* 31.4 (1966), pp. 581–603.

[122] Abdolreza Abhari and Mojgan Soraya. "Workload generation for YouTube." In: *Multimedia Tools Appl.* 46.1 (Jan. 2010), pp. 91–118.

[123] Xu Cheng, Cameron Dale, and Jiangchuan Liu. "Understanding the Characteristics of Internet Short Video Sharing: YouTube as a Case Study." In: *CoRR* abs/0707.3670 (2007).

[124] David Hadaller et al. "Vehicular opportunistic communication under the microscope." In: *Proceedings of the 5th international conference on Mobile systems, applications and services*. MobiSys '07. San Juan, Puerto Rico: ACM, 2007, pp. 206–219.

[125] Thiago H. Silva et al. "A Comparison of Foursquare and Instagram to the Study of City Dynamics and Urban Social Behavior." In: *Proceedings of the 2Nd ACM SIGKDD International Workshop on Urban Computing*. UrbComp '13. Chicago, Illinois: ACM, 2013, 4:1–4:8.

[126] T.H. Silva et al. "Challenges and opportunities on the large scale study of city dynamics using participatory sensing." In: *Computers and Communications (ISCC), 2013 IEEE Symposium on*. July 2013, pp. 000528–000534.

[127] Ling-Yin Wei, Yu Zheng, and Wen-Chih Peng. "Constructing Popular Routes from Uncertain Trajectories." In: *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '12. Beijing, China: ACM, 2012, pp. 195–203.

[128] *Mobile context-Adaptive CAching for COntent-centric networking*. URL: https://macaco.inria.fr.

[129] *Priva'Mov - Mobilité et vie privée*. URL: http://liris.cnrs.fr/privamov/.